Gestaltung eines Rezept-Recommender-Systems

Bachelorarbeit

Hochschule der Medien Stuttgart
Fachbereich Information und Kommunikation
Studiengang Wirtschaftsinformatik und digitale Medien

Prof. Dr. Hendrik Meth
Prof. Dr. Martin Engstler

Sommersemester 2020

Vorgelegt von: Daniel Volz
Matrikelnummer: 29558

Stuttgart, 01.06.2020

HOCHSCHULE DER MEDIEN

Ehrenwörtliche Erklärung

Name:	Volz	Vorname:	Daniel
Matrikel-Nr.:	29558	Studiengang:	Wirtschaftsinformatik und digitale Medien

Hiermit versichere ich, Daniel Volz ehrenwörtlich, dass ich die vorliegende Bachelorarbeit mit dem Titel: „Gestaltung eines Rezept-Recommender-Systems" selbständig und ohne fremde Hilfe verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen wurden, sind in jedem Fall unter Angabe der Quelle kenntlich gemacht. Die Arbeit ist noch nicht veröffentlicht oder in anderer Form als Prüfungsleistung vorgelegt worden.

Ich habe die Bedeutung der ehrenwörtlichen Versicherung und die prüfungsrechtlichen Folgen (§ 26 Abs. 2 der Bachelor-SPO (6-Semester), § 24 Abs. 2 Bachelor-SPO (7-Semester), § 23 Abs. 2 Master-SPO (3 Semester) bzw. §19 Abs. 2 Master-SPO (4 Semester und berufsbegleitend) der HdM) einer unrichtigen oder unvollständigen ehrenwörtlichen Versicherung zur Kenntnis genommen.

Böblingen, 01.06.2020 .....

Ort, Datum

.....

Unterschrift

Kurzfassung

Die vorliegende Arbeit befasst sich mit dem Thema der Berechnung von Ähnlichkeiten zwischen Kochrezepten. Dabei werden zum einen Ähnlichkeiten auf Zutatenbasis und zum anderen auf Nährwertbasis zwischen Kochrezepten berechnet. Für die Berechnung der Zutatenähnlichkeit wird die Jaccard Distanz benutzt. Die Nährwert-Ähnlichkeit wird mit der Methode der Euklidischen Distanz berechnet. Als Datenbasis dienen 6300 englischsprachige Kochrezepte der Kategorie Hauptgericht der Webseite Allrecipes.com.

Ziel dieser Arbeit war es, eine Methode zu entwickeln, die im Sinne eines Rezept-Recommender-Systems, einem Endnutzer zunächst Rezepte empfiehlt, die Ähnlichkeiten zu seiner persönlichen Kochrezeptsammlung aufweisen. In einem weiteren Schritt sollen dem Endnutzer Kochrezepte hinsichtlich ihrer Nährwerteigenschaften empfohlen werden. Die empfohlenen Kochrezepte sollen Nährwerte aufweisen, die dem empfohlenen Tagesbedarf für eine Hauptmahlzeit entsprechen und zusätzlich niedrig an Fett- oder niedrig an Kohlenhydratgehalt sein. Drei unterschiedliche Methoden zur Berechnung der Zutaten-Ähnlichkeit werden am Ende dieser Arbeit miteinander verglichen und rein zufälligen Werten gegenübergestellt.

Schlagworte: Recommender-System, Kochrezept, Ähnlichkeit, Jaccard Distanz, Euklidische Distanz

Abstract

This thesis deals with the topic of calculating similarities between recipes. On the one hand, similarities are calculated on the basis of ingredients and on the other hand on the nutritional value of recipes. The Jaccard distance is used to calculate the similarity between the ingredients. The nutritional similarity is calculated with the Euclidean distance method. The database is based on 6300 English recipes of the category main dish from the website Allrecipes.com.

The aim of this thesis was to develop a method which – in the sense of a recipe-recommender-system – in a first step recommends recipes to an end user which show similarities to his/her personal recipe collection. In a further step, the end user will be recommended recipes regarding their nutritional properties. The recommended recipes should have nutritional values that correspond to the recommended daily requirement for a main meal and in addition be low in fat or low in carbohydrates. Three different methods for calculating the similarity of ingredients have been evaluated and compared with each other and with random values.

Keywords: Recommender-system, recipe, similarity, Jaccard distance, Euclidean distance

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

DGE	Deutsche Gesellschaft für Ernährung e.V.
FAO	Food and Agriculture Organization of the United Nations
HdM	Hochschule der Medien, Stuttgart
IQR	Interquartile range
JSON	JavaScript Object Notation
NCBI	National Center for Biotechnology Information
OECD	Organization for Economic Co-operation and Development

1 Einführung

In der vorliegenden Arbeit soll ein Modell entwickelt werden, dass das Vorschlagen von Kochrezepten hinsichtlich ihrer Ähnlichkeit ermöglicht. Gegenwärtig basieren die meisten Ansätze zum Vorschlagen von Kochrezepten auf dem Konzept der Recommender Systeme, die mit Hilfe von machine-learning umgesetzt werden. Die Modelle, die in dieser Arbeit vorgestellt werden basieren auf Ähnlichkeitsmaßen, welche die Ähnlichkeiten zwischen Rezepten berechnen. Als Datenbasis für das Testen der Modelle, werden Rezeptdaten der Webseite Allrecipe.com verwendet.

1.1 Problemstellung und Themenrelevanz

Ernährung ist ein Thema, mit dem sich jeder Mensch zwangsläufig jeden Tag auseinandersetzen muss. So trivial diese Aussage für sich ist, umso komplexer wird der Zusammenhang, wenn man sich die Probleme anschaut, die auf globaler Ebene mit dem Thema Nahrung verbunden sind. Eines der akuten Probleme, zumindest in der westlichen Welt und zunehmend auch im asiatischen Raum, sind sogenannte Zivilisationskrankheiten, die durch Mangel und Fehlernährungen ausgelöst werden. Eine in Industrie- und Schwellenländern weit verbreitete Zivilisationskrankheit ist Adipositas (krankhaftes Übergewicht). Laut der *OECD*-Studie *The Heavy Burden of Obesity* aus dem Jahr 2019, werden in den nächsten 30 Jahren geschätzt 92 Millionen Menschen der *OECD*-Länder an den Folgen von Adipositas und den damit verbundenen Krankheiten sterben (vgl. Devaux & Vuik, 2019, S. 14). Um der Entwicklung von Mangel- und Fehlernährung entgegenzuwirken bieten automatische Auswertungen und Interpretationen von Ernährungsgewohnheiten möglicherweise ein geeignetes Mittel. Es ist vorstellbar, dass geeignete Modelle die individuellen Ernährungsgewohnheiten eines Menschen auswerten, um ihn auf problematische Ernährungsgewohnheiten aufmerksam zu machen und ihm bessere Alternativen aufzuzeigen. Dabei geht es aber nicht nur um offensichtliche Fehlernährung, wie das einseitige Konsumieren von nährstoffarmer und stark kalorienhaltiger Nahrung,

sondern vor allem um subtilere Formen der Fehlernährung, wie z. B. Vitamin- und Mineralstoffmangel.

Eine mögliche Anwendung solcher Modelle wäre z. B. eine Smartphone Applikation, die die Essensgewohnheiten des Nutzers ermittelt und ihm nicht nur neue Rezepte vorschlägt, sondern auch darauf achtet, dass die vorgeschlagenen Rezepte individuell auf seinen Nährstoffhaushalt abgestimmt sind. Ein weiteres mögliches Anwendungsgebiet für das Auswerten und Interpretieren von Ernährungsgewohnheiten liegt im Bereich der Beeinflussung und Veränderung von individuellem Ernährungsverhalten. So könnten den Menschen beispielsweise mögliche Alternativen zum Fleischkonsum aufgezeigt werden. Laut einer Studie der *Ernährungs- und Landwirtschaftsorganisation der Vereinten Nationen* (FAO) aus dem Jahr 2006, ist der Fleischkonsum und die damit verbundenen landwirtschaftlichen Aktivitäten für 18% der weltweiten Klimagas-Emissionen verantwortlich (vgl. FAO, 2006, S. 112). Um diesem Trend entgegenzuwirken ist es vorstellbar ein Modell zu entwickeln, das dem Nutzer fleischlose Alternativen auf der Basis seiner Essgewohnheiten vorschlägt. Somit könnte eine Reduktion von tierischen Produkten in der täglichen Ernährung nicht als etwas Negatives oder als ein Mangel wahrgenommen werden, sondern als eine neue Möglichkeit seine Lieblingszutaten zu einer fleischlosen Alternative zu verbinden.

Da die Ernährung ein zentraler Baustein der Gesundheit des Menschen ist, hat sie auch großes Potential die Lebensqualität eines Menschen zu verbessern. Dieses Potential kann womöglich mit neuen Methoden der Datenverarbeitung und Auswertung ausgeschöpft werden.

1.2 Zielsetzung

Die Kernidee dieser Arbeit besteht im ersten Schritt darin Ähnlichkeiten zwischen zwei oder meist mehreren Kochrezepten zu berechnen, um daraus Rezeptvorschläge zu erzeugen. Für einen Rezeptvorschlag benötigt man zentral zwei Bausteine, zum einen ein Benutzerprofil, in dem mehrere favorisierte Rezepte eines Benutzers gespeichert sind und zum anderen eine Datenbank mit Rezepten, die mit den Rezepten des Benutzerprofils verglichen werden. Das Ergebnis dieses Vergleichs ist z. B. eine Top 10 Liste mit Rezepten

aus der Datenbank, die absteigend nach ihrem Ähnlichkeitswert sortiert sind. Das bedeutet, dass das Rezept mit der höchsten Ähnlichkeit zum Benutzerprofil sich an erster Stelle der Top 10 Liste befindet. Im zweiten Schritt wird eine zweite Ähnlichkeitsberechnung auf der Basis der Nährwertinformationen aller Rezepte der Top 10 Liste durchgeführt. Dabei dienen nicht mehr die Zutaten eines Rezepts als ein Kriterium für die Ähnlichkeit, sondern die individuellen Nährwerte eines Rezepts wie z. B. Fett, Kohlenhydrate und Eiweiß. Das Ziel des zweiten Schritts ist es wiederum eine Top 10 Liste zu erstellen, die nicht nur nach Zutatenähnlichkeit, sondern nach zuvor festgelegten Nährwertkriterien wie z. B. einem niedrigen Fettgehalt und einem hohen Eiweißgehalt berechnet ist. Das gewünschte Resultat sollte eine Liste mit Rezepten sein, die zum einen viele gleiche Zutaten des Benutzerprofils beinhaltet und zum anderen zusätzlich die gewünschten Nährwerte (z. B. low-fat) haben.

Zur Berechnung der Ähnlichkeitswerte werden zwei mathematische Verfahren angewandt. Diese Verfahren sind der Jaccard Koeffizient und die Euklidische Distanz. Der Jaccard-Koeffizient wird für die Berechnung der Zutaten-Ähnlichkeit angewandt und für die Berechnung der Nährwert-Ähnlichkeit wird die Euklidische Distanz eingesetzt.

Das primäre Ziel dieser Arbeit ist es ein hybrides Modell zu entwickeln, das im ersten Schritt die Zutaten-Ähnlichkeit von einem Benutzer-Profil zu der Rezept-Datenbank berechnet. Im zweiten Schritt werden die ähnlichsten Rezepte als Grundlage genutzt, um Nähwert-Ähnlichkeiten für die Nährwert-Profile low-fat und low-carb zu berechnen. Das hybride Modell ist somit in zwei Teile aufgeteilt:

1. *Zutaten-Ähnlichkeit*: Zur Berechnung der Zutaten-Ähnlichkeit werden die drei unterschiedlichen Methoden Naive Berechnung, Kürzeste Distanz Berechnung und die Rezept Vektor Berechnung vorgestellt. Diese drei Methoden werden miteinander verglichen und nach zuvor festgelegten Kriterien bewertet. Die Ergebnisse werden in Kapitel 6.3 diskutiert.

2. *Nährwert-Ähnlichkeit*: Die Nährwert-Ähnlichkeit soll dazu dienen dem Benutzer Rezeptvorschläge zu generieren, die als gesund definiert sind. Als gesund wird hier der empfohlene Tagesbedarf an Nährstoffen pro Hauptmahlzeit

definiert. Auf der Grundlage des empfohlenen Tagesbedarfs an Nährwerten werden zwei Profile low-fat und low-carb erstellt. Die Makronährwerte dieser Profile werden hinsichtlich einer fettreduzierten und kohlenhydratarmen Ernährungsweise angepasst, um dem Benutzer diätische und gesunde Rezepte zu empfehlen. In Kapitel 6.4 werden die Nährwerte der empfohlenen Rezepte mit den Rezepten des Benutzer-Profils verglichen, um festzustellen ob die generierten Rezepte tatsächlich die diätischen Zielnährwerte erreichen.

Das sekundäre Ziel der Arbeit ist das Ausarbeiten und Aufzeigen von notwendigen Einzelschritten, die im Bereich der Datenanalyse anfallen und für die Modellbildung wichtig sind. Zu diesen Schritten gehören die Auswahl der Datenquelle, das Auslesen der Daten, die Säuberung und Aufbereitung der gesammelten Daten, die Auswertung der Daten mit geeigneten Werkzeugen und schließlich die Visualisierung und Interpretation der Ergebnisse basierend auf dem oben beschriebenen Modell.

Im Detail soll das primäre Ziel dieser Arbeit folgendermaßen verfolgt werden:

Das hybride Modell wird an zwei Benutzerprofilen getestet. Das erste Benutzer-Profil ist das *Fett-Profil* und das zweite Profil ist das *Normal-Profil*. Das Fett Profil enthält Rezepte, die einen überdurchschnittlichen Anteil an Fettgehalt haben. Für dieses Benutzer-Profil wird das low-fat-Profil verwendet, um dem Nutzer spezielle Rezepte zu empfehlen, die wenig Fett haben. Das Ziel hierbei ist es dem Nutzer nicht nur low-fat Rezepte vorzuschlagen, sondern die Rezepte, die zu seinem Benutzer-Profil passen, das heißt eine hohe Zutaten-Ähnlichkeit besitzen.

Das Normal-Profil enthält Rezepte mit zufälligen Nährwertangaben. Das Profil dieser Benutzter soll keine Extreme, sondern eher den durchschnittlichen Nutzer darstellen. Dem durchschnittlichen Nutzer sollen mit Hilfe des low-carb Profils Rezepte vorgeschlagen werden, die einen niedrigen Kohlenhydratwert besitzen.

Schließlich soll die entwickelte Methodik an drei Hypothesen getestet werden. Die ersten beiden Hypothesen beziehen sich auf die Zutaten- und Nährwert-Ähnlichkeit. Ein mögliches Kriterium für die Güte der Zutaten-Ähnlichkeit ist z. B. die Überschneidung an

Zutaten, die sowohl im Profil des Nutzers vorkommen als auch im Rezeptvorschlag der hybriden Methode. Je mehr Zutaten der Rezepte eines Benutzer-Profils in den Rezeptempfehlungen vorkommen, desto besser ist die Empfehlung. Die erste Hypothese lautet:

Hypothese 1: *„Wenn ein Rezeptvorschlag mit der Jaccard Distanz generiert wird, dann hat er eine höhere Überschneidung an Zutaten zum Benutzter-Profil als ein zufällig generierter Rezeptvorschlag.“*

Die zweite Hypothese bezieht sich auf die Berechnung der Nährwert-Ähnlichkeit und lautet:

Hypothese 2: *„Wenn ein Rezeptvorschlag mit der Euklidischen Distanz berechnet wird, dann sind die Nährwerte der Rezepte näher an den Nährwerten des low-fat bzw. low-carb Profils als ein zufällig generierter Rezeptvorschlag.“*

Die dritte Hypothese bezieht sich auf die Anzahl der Rezepte im Benutzter-Profil. Man kann annehmen, dass die Güte eines Rezeptvorschlags von der Menge an Rezepten im Benutzter-Profil abhängt. Die dazugehörige Hypothese lautet:

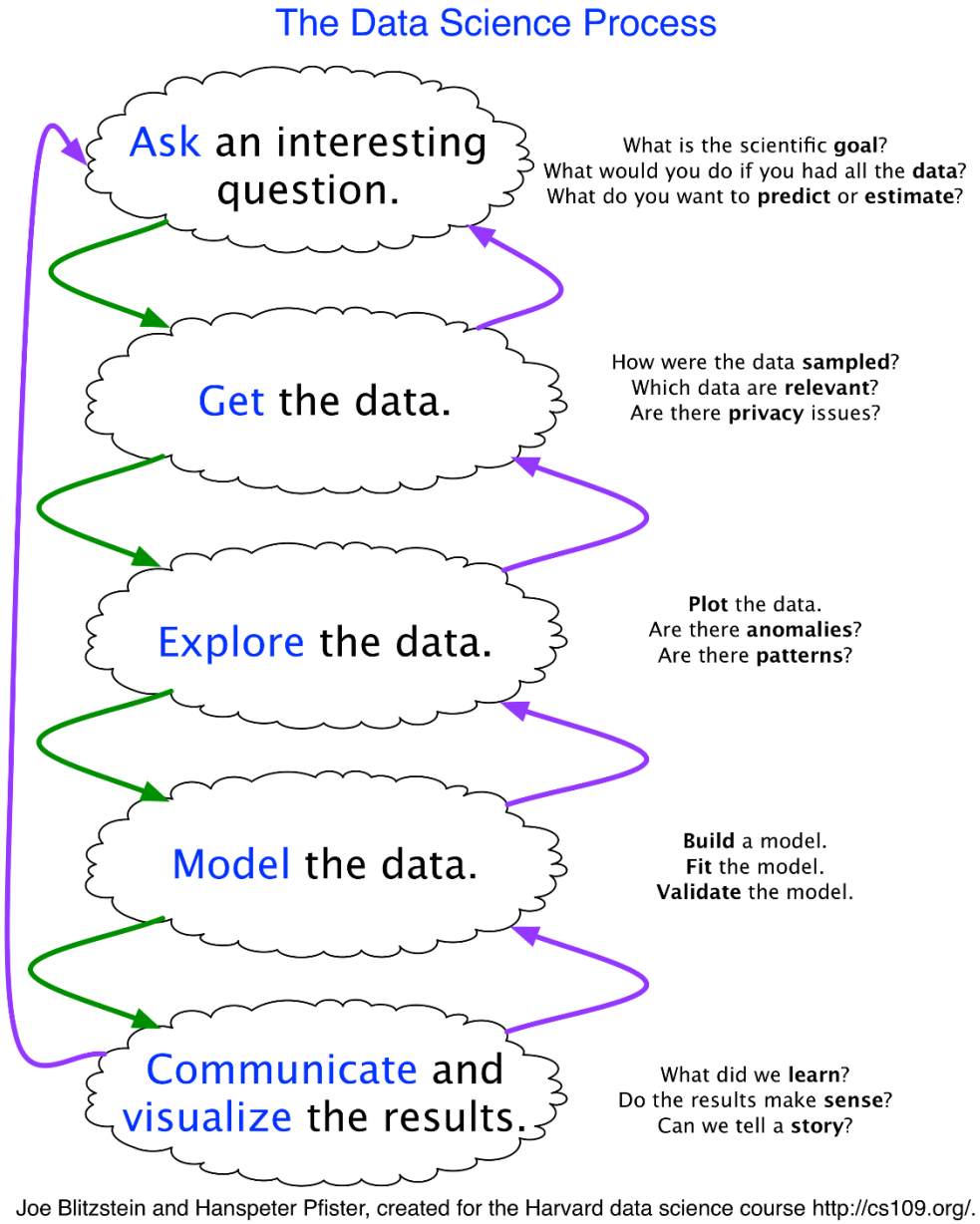
Hypothese 3: *„Je mehr Rezepte ein Benutzer in seinem Profil besitzt, desto höher ist die Überschneidung zwischen den Benutzter-Profil-Zutaten und den Rezeptvorschlags-Zutaten.“*

Diese drei Hypothesen gilt es im Schlussteil dieser Arbeit zu überprüfen.

1.3 Aufbau der Arbeit und Forschungsansatz

Der Aufbau dieser Arbeit orientiert sich an dem Data-Science-Process-Framework. Das Framework wurde im Rahmen einer Computer-Science-Vorlesung der Universität Harvard von den beiden Dozenten Pfister und Blitzstein vorgestellt. Der Data-Science-Prozess ist ein iterativer und nicht-linear Prozess, der aus fünf grundsätzlichen Phasen besteht (vgl. Blitzstein & Pfister, 2013).

Abbildung 1 Framework Data-Science-Process



Die fünf Phasen des Data-Science-Prozesses

(Quelle: vgl. Blitzstein & Pfister, 2013)

1. Fragestellung
2. Datenakquise
3. Datenexploration
4. Modellbildung
5. Kommunikation der Ergebnisse

Angelehnt an die fünf Phasen des Data-Science-Prozesses ist diese Arbeit folgendermaßen strukturiert. Im ersten Kapitel wurden bereits die Ziele der Arbeit und die daraus abgeleiteten Hypothesen vorgestellt. Im zweiten und dritten Kapitel wird die gegenwärtige Forschung zum Thema Kochrezepte und Empfehlungssysteme vorgestellt sowie die mathematischen Verfahren, die in dieser Arbeit zum Einsatz kommen beschrieben. Die Themenfelder Datenakquise, Datenexploration und Datenstruktur werden im vierten Kapitel behandelt. Im fünften Kapitel wird das hybride Modell für die Berechnung der Zutaten- und die Nährwert-Ähnlichkeit erörtert. Zum Ende der Arbeit werden in den letzten beiden Kapiteln die Ergebnisse vorgestellt und interpretiert.

2 Stand der Forschung

In diesem Kapitel werden verschiedene Positionen zum Thema Recommender Systeme für Kochrezepte vorgestellt, die sich auf unterschiedliche Art und Weise mit dem Thema auseinandergesetzt haben.

In ihrer Arbeit *Recommending Food: Reasoning on Recipes and Ingredients* vergleichen Freyne und Berkovsky verschiedenen Filtermethoden wie content-based, collaborative und hybrides filtering hinsichtlich ihrer Performanz bei der Empfehlung von Kochrezepten. Grundlage der Empfehlungen sind dabei Rezeptbewertungen von 512 Nutzern. Diese Bewertungen wurden nicht nur auf Rezepte angewendet, sondern eins-zu-eins auf die in den Rezepten vorhanden Zutaten übertragen. Um dem Anwender neue Rezepte vorzuschlagen wurden Rezepte gesucht, die die größte Gemeinsamkeit hinsichtlich ihrer Zutaten und den präferierten Zutaten des Anwenders besitzen.

Beim Vergleich der verschiedenen Filtermethoden hat sich gezeigt, dass eine hybride Methode aus content-based und collaborative filtering die besten Ergebnisse erzielt hat (vgl. Freyne & Berkovsky, 2010, S. 4ff.).

Die Frage, ob die Präferenzen eines Nutzers hinsichtlich seiner Rezeptwahl vorhersagbar sind und welche Schlüsselfaktoren die Vorhersage beeinflussen behandeln Teng, Lin und Adamic in ihrer Arbeit mit dem Titel *Recipe Recommendation using Ingredient Networks*. Primär wird in diesem Artikel diskutiert, ob einzelne Zutaten in einem Gericht essenziell für das Rezept sind oder weggelassen bzw. in ihrer Menge verändert werden können.

Als zentrales methodisches Element wurden zwei unterschiedliche Netzwerktypen erstellt, die die Beziehung zwischen Zutaten beschreiben. Das complement-network erfasst, welche Zutaten dazu tendieren häufig miteinander aufzutreten und zu welcher Geschmacksgruppe (süß oder salzig) sie gehören. Das substitute-network enthält Informationen hinsichtlich der Ersetzbarkeit von Zutaten bezogen auf gesundheitliche Faktoren. Diese Informationen wurden anhand von Nutzerkommentaren im Kommentarbereich der jeweiligen Rezepte gewonnen.

Teng, Lin und Adamic zeigen in ihrer Arbeit, dass Vorhersagen von Rezeptvorschlägen vor allem dann eine hohe Genauigkeit erzielen, wenn sie auf dem complement-network basiert. Das heißt wenn nicht die gesamte Zutatenliste eines Rezepts verwendet wird, sondern nur die Informationen welche Zutaten häufig in Kombination miteinander auftreten (vgl. Teng et al., 2011, S. 8ff.).

Ahn et al. gehen in ihrer Arbeit *Flavor Network and the Principles of Food Pairing* der Frage nach, ob es allgemeine Muster in der Kombination von Zutaten und Geschmäckern gibt. Mit einem Netzwerk-basierten Ansatz untersuchen sie den Einfluss von Geschmacksverbindungen auf unterschiedliche Zutatenkombinationen. Durch die chemische Analyse von Zutaten identifizieren sie, dass jede Zutat im Durchschnitt 51 Geschmackverbindungen aufweist. Mithilfe der Geschmacksverbindungen erstellen sie ein sogenanntes Geschmacks-Netzwerk (flavor network). Basierend auf den gewonnenen Erkenntnissen können Ahn et al. eine Serie von statistisch signifikanten Mustern erkennen, die charakteristisch für die Auswahl und Kombination von Zutaten in menschengemachten Gerichten sind. Diese Muster lassen sich in verschiedene geographische Regionen einteilen. Schließlich kommen Ahn et al. zu dem Ergebnis, dass nordamerikansiche und europäische Gerichte jene Zutaten aufweisen, die ähnliche oder gleiche Geschmacksverbindungen aufweisen, wohingegen asiatische Gerichte das nicht tun (vgl. Ahn et al., 2011, S. 5f.).

Die Arbeit *Computational Creativity in the Culinary Arts* von Cromwell, Galeota-Sprung und Ramanujan behandelt das Thema Rezeptempfehlung aus einer gänzlich anderen Perspektive. Zwar befasst sich ihre Arbeit mit der künstlichen Generierung von neuartigen Salatrezepten, auf der Grundlage von gut bewerteten *echten* (von Menschen verfassten) Salatrezepten. Für die künstliche Rezeptgenerierung wurde zunächst eine Klassifizierungsmethode erarbeitet, welche anhand von Rezeptzutaten, das besser passende Rezept aus zwei Rezepten auswählte. Für die Verbesserung der Rezeptqualität wurde zusätzlich das Konzept des complement-Netzwerk von Teng, Lin, und Adamic sowie die Geschmacks-Verbindungs-Daten von Ahn et al. eingesetzt.

Für die Rezeptbewertung wurde eine Ranking-Methode erarbeitet, die die generierten Rezepte mit bereits bekannten und als gut bewerteten Rezepten verglich und diese anhand eines Zahlenwerts (score) als *besser* oder als *schlechter* kategorisierte. Nach dem Durchlauf von mehreren Iterationen wurden die Top 20 künstlichen Rezepte ermittelt. In einer Blindverkostung wurden drei generierte Rezepte mit drei echten Rezepten verglichen. Im Durschnitt schnitten alle drei generieten Rezepte in der Geschmacksbewertung schlechter ab als ihre echten Pendants. Bei der Bewertung der Neuartigkeit hinsichtlich der verwendeten Zutaten erzielten die generierten Rezepte im Durchschnitt die besseren Bewertungen (vgl. Cromwell et al., 2015, S. 39ff.).

3 Theoretische Grundlagen

In diesem Kapitel werden die 3 klassischen Ansätze für Empfehlungssysteme beschrieben. Das Empfehlungssystem, welches in dieser Arbeit verwendet wird ist ausschließlich das content-based filtering. Die anderen beiden Systeme sind hier der Vollständigkeitshalber vorgestellt.

Am Ende des Kapitels werdend die Grundlagen der beiden Distanzmaße Jaccard Distanz und Euklidische Distanz erörtert.

3.1 Empfehlungssysteme (Recommender Systems)

Empfehlungssysteme (Recommender Systems) sind Software-Programme oder auch Algorithmen, die einem Nutzer, basierend auf seinen Verhalten, neue Produktempfehlungen generieren. Ein Empfehlungssystem kann auch als ein Informationsfilter verstanden werden, der aus einer großen Menge an Informationen (z. B. Bücher, Musik oder Filme) nur die für den Nutzer relevanten Inhalte herausfiltert (vgl. Portugal et al., 2015, S. 2).

Eines der bekanntesten Beispiele für Empfehlungssysteme ist die Website Amazon.de. Amazon nutzt unterschiedliche Empfehlungsstrategien, um ihren Kunden individuelle Kaufanreize zu ermöglichen. Diese Strategien sind im Detail nicht öffentlich einsehbar, jedoch lässt sich beobachten, dass Amazon zumindest die drei klassischen Recommender-Systeme content-based filtering, collaborative filtering und hybrides filtering für ihre Kaufempfehlungen benutzt. Diese drei Empfehlungssysteme werden im Folgenden beschrieben.

3.1.1 Content-based filtering

Die erste Empfehlungsstrategie lässt sich mit dem Satz: ‚Das könnte Ihnen gefallen' zusammenfassen. Dabei werden den Kunden Produkte empfohlen, die entweder ähnlich zu den Produkten sind, die sie bereits in der Vergangenheit gekauft haben oder jene Produkte, die sie sich in der Vergangenheit angeschaut aber nicht gekauft haben. Grundsätzlich lässt sich diese Empfehlungsstrategie auch ohne Kundenkonto anwenden, da die

Ähnlichkeiten nur zwischen den Items bzw. deren Eigenschaften berechnet werden. Dieses Verfahren nennt sich *content-based filtering*, da für die generierten Empfehlungen der bereits vorhandene Inhalt (*content*) wie z. B. Artikel und die Artikeleigenschaften genutzt werden (vgl. Pazzani & Billsus, 2006, S. 329).

3.1.2 Collaborative filtering

Die zweite Strategie lässt sich mit dem Satz: ‚Kunden, die diesen Artikel angesehen haben, haben auch folgende Artikel angesehen' zusammenfassen. Wie man bereits dem Satz entnehmen kann, handelt es sich bei dieser Methode um eine Empfehlung, die auf den Informationen von anderen Nutzern basiert. Dieses Verfahren nennt sich *collaborative-filtering.* Bei diesem Verfahren vergleicht man Kundeninformationen wie z. B. besuchte Artikelseiten zwischen zwei Kunden. Wenn zwei Kundenprofile sich ähnlich sind, so empfiehlt das Recommender System dem Kunden A die Artikel, die Kunde B in der Vergangenheit gekauft hat. Die Idee dahinter ist, dass Kundenprofile mit ähnlichen Interessen auch an den ähnlichen Produkten interessiert sind (vgl. ebd., S. 339).

3.1.3 Hybrides filtering

Die dritte Empfehlungsstrategie ist eine Mischung aus collaborative filtering und content-based filtering. So können beispielsweise im ersten Schritt, mit Hilfe des collaborative filterings, zunächst ähnliche Kundenprofile gesucht werden. Ähnlichkeitskriterien können demographische Merkmale wie Alter, Wohnort und Geschlecht sein. Im zweiten Schritt kann anschließend mit einen content-based filtering nach ähnlichen Produktinteressen gefiltert werden, um den Kunden neue Kaufvorschläge zu empfehlen (vgl. Portugal et al., 2015, S. 3).

3.2 Jaccard-Koeffizient

Der Jaccard-Koeffizient (J) oder auch Jaccard-Index ist ein statistisches Mittel, um die Ähnlichkeit oder Unähnlichkeit zwischen zwei Mengen zu berechnen. Um den Jaccard-Koeffizienten zu berechnen, muss man die Anzahl der gleichen Ausprägungen

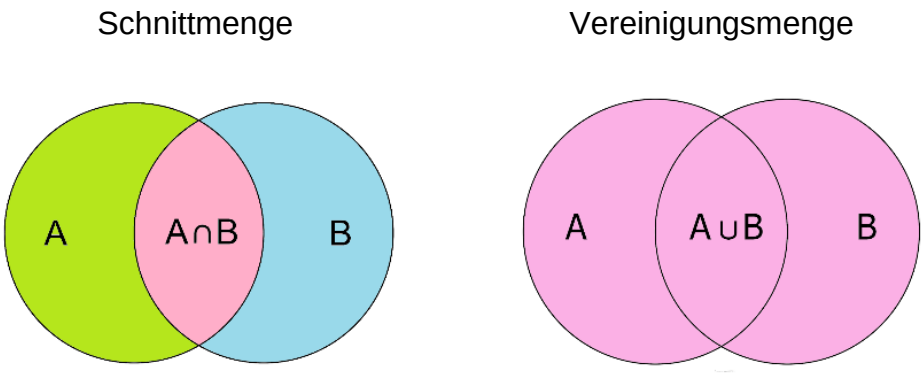
der Schnittmenge eines Datensets durch die Vereinigungsmenge des Datensets A und B teilen. Die Formel für die Berechnung des Jaccard-Koeffizienten ist in (1) abgebildet.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (1)$$

Die Jaccard Distanz (d_J) oder auch Jaccard-Metrik lässt sich aus dem Jaccard-Koeffizienten ableiten. Wie in (2) dargestellt, berechnet sich die Jaccard Distanz indem man die Zahl 1 von dem Jaccard-Koeffizienten abzieht. Oder man berechnet sie indem man die Vereinigungsmenge von der Schnittmenge abzieht und das Ergebnis durch die Vereinigungsmenge der Datensets A und B teilt (vgl. Niwattanakul et al., 2013, S. 2).

$$d_J(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \quad (2)$$

Abbildung 2 Beispiel für die Vereinigung von Datenset A und B



(Quelle: Eigene Darstellung)

Ein Beispiel für die Berechnung der Jaccard Distanz zwischen zwei Rezepten kann wie folgt berechnet werden. Gegeben sind zwei Rezepte mit den folgenden Zutaten-IDs:

1. Rezept 1 hat die Zutaten-IDs: 1, 3, 4, 6, 8.
2. Rezept 2 hat die Zutaten-IDs: 3, 4, 5, 6, 8.

Setzt man die Mengen für die Zutaten-IDs der beiden Rezepte in die Formel (3) für die Berechnung der Jaccard Distanz (d_J) ein, so erhält man für die Distanz zwischen Rezept 1 und Rezept 2 einen Wert von $0,\overline{4}$.

$$d_J(R_1, R_2) = 1 - \frac{|R_1 \cap R_2|}{|R_1 \cup R_2|} = 1 - \frac{|\{3,4,6,8\}|}{|\{1,3,4,5,6,8\}|} = 1 - \frac{4}{6} = 1 - 0,\overline{6} = \underline{0,\overline{4}} \quad (3)$$

3.3 Euklidische Distanz

Die Euklidische Distanz ist ein Abstandmaß das Entfernungen zwischen zwei Punkten in der Ebene oder im Raum messen kann. Der Abstand (d) zwischen zwei Punkten p und q in einem n-dimensionalen Raum lässt sich wie in (4) dargestellt berechnen (vgl. Huang, 2008, S. 51).

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2} = \quad (4)$$

$$\sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

In dieser Arbeit wird die Euklidische Distanz angewendet, um die Entfernung zwischen den Nährwertinformationen von Kochrezepten zu berechnen. Dabei wird z. B. zunächst ein Referenzpunkt p_{ref} definiert, der die empfohlenen Nährwertinformationen für eine Mahlzeit pro Tag enthält. Da es pro Rezept 20 Nährwertinformationen gibt, befindet sich der Punkt p_{ref} in einem 20-dimensionalen Raum. Nun kann man mit Hilfe der Euklidischen Distanz den Abstand zu den Nährwertinformationen eines jeden anderen Rezeptes bzw. eines Punktes q_{Rezept} berechnen. Je kleiner dabei der Abstand $d(p_{ref}, q_{rezept})$ ausfällt, desto ähnlicher sind die Nährwerte eines Rezeptes zu den Nährwerten des Referenzrezeptes.

Ein Beispiel für die Berechnung der Distanz zwischen den Nährwertinformationen zweier Rezepte kann wie folgt aussehen (die verwendeten Nähwertangaben sind fiktiv und zur übersichtlichen Darstellung wurden nur 3 statt 20 Nährwerte verwendet):

1. Das Referenzrezept p_{ref} hat die Makronährwerte:
   Fett: 20g, Kohlenhydrate: 100g, Eiweiß: 40g.
2. Das Datenbank-Rezept q_{rezept} hat die Makronährwerte:
   Fett: 50g, Kohlenhydrate: 100g, Eiweiß: 10g.

$$d(p_{ref}, q_{rezept}) = \sqrt{(20-50)^2 + (100-100)^2 + (40-10)^2} \quad (5)$$
$$= \sqrt{(-30)^2 + (0)^2 + (30)^2} = \sqrt{1800} = \underline{42{,}426}$$

Die Distanz der Nährwertinformationen zwischen Referenzrezept p_{ref} und Datenbank-Rezept q_{rezept} beträgt 42, 426.

4 Datenbasis

Dieses Kapitel befasst sich mit der Datenbasis, die für diese Arbeit notwendig war. Zunächst wird das Thema Datenquelle und Datenakquise besprochen. Im Anschluss wird die Datenqualität diskutiert und die Datenstruktur beschrieben. Zum Schluss wird gezeigt, wie die Daten für den Import in eine Programmierumgebung modifiziert und letztlich für die Berechnungen bereinigt wurden.

Die Rezeptdaten, die für diese Arbeit benötigt werden, wurden von der Webseite Allrecipes.com mit der Hilfe eines eigens dafür erstellten Webcrawlers gesammelt. Ein Beispiel dafür wie ein zufälliges Rezept auf der Webseite aussieht zeigt Abbildung 3. Allrecipes.com ist laut eigenen Angaben, mit seinen monatlich 30 Millionen einzelnen Besuchern (unique visitors), das größte auf Nahrung spezialisierte soziale Netzwerk der Welt. Allrecipes.com hatte im Jahr 2015 zehn Millionen Benutzter, die jährlich 55 Millionen auf Nahrungsmittel fokussierte Inhalte erstellt, gespeichert und geteilt haben (vgl. Allrecipes.com, 2015).

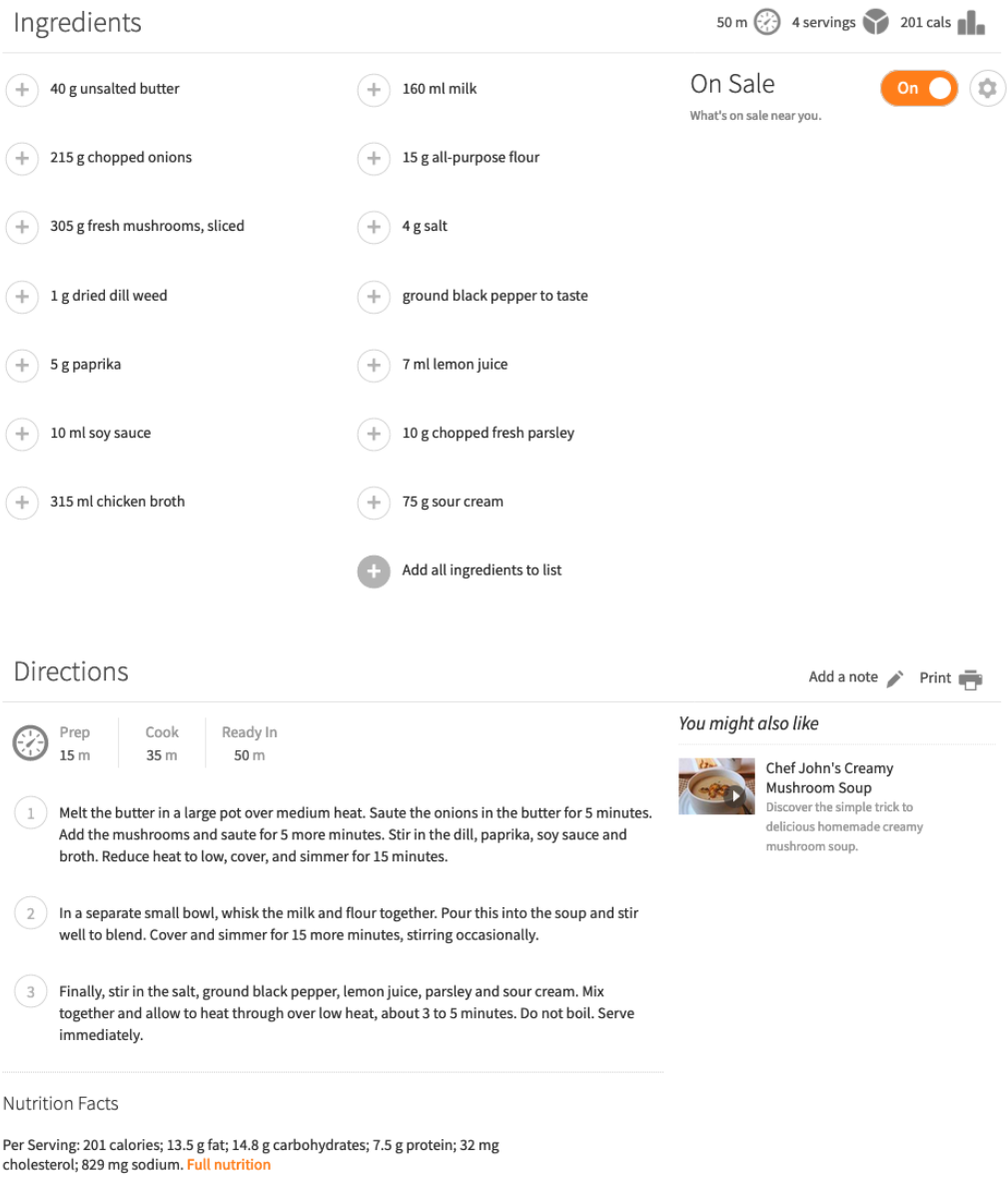
Neben seiner großen Internet-Reichweite und der großen Anzahl an Rezepten, spricht auch der hohe Informationsgehalt eines jeden Rezeptes für die Webseite Allrecipes.com als Datenquelle. Neben den üblichen Angaben wie Zutaten, Menge, Kochdauer, Schwierigkeitsgrad usw. bietet die Webseite zusätzlich detaillierte Nähstoffangeben für jedes seiner Rezepte an.

4.1 Datenakquise und Datenbeschreibung

Mit Hilfe des open-source web-crawling-frameworks Scrapy wurde ein Bot-Programm erstellt, welches 6305 Rezepte der Kategorie Hauptgericht aus der Webseite Allrecipes.com in eine MongoDB Datenbank abspeichert. In dieser Arbeit werden bewusst alle anderen Kategorien wie z. B. Frühstück, Dessert usw. ausgeblendet, um den Aufwand einzugrenzen und eine Vergleichbarkeit zwischen den Rezepten zu ermöglichen.

MongoDB Datenbanken sind dokumentbasierte Datenbanken, die vor allem für das Speichern von JSON-Objekten verwendet werden. Ein Beispiel für ein als JSON-Objekt gespeichertes Rezept ist in Abbildung 4 dargestellt.

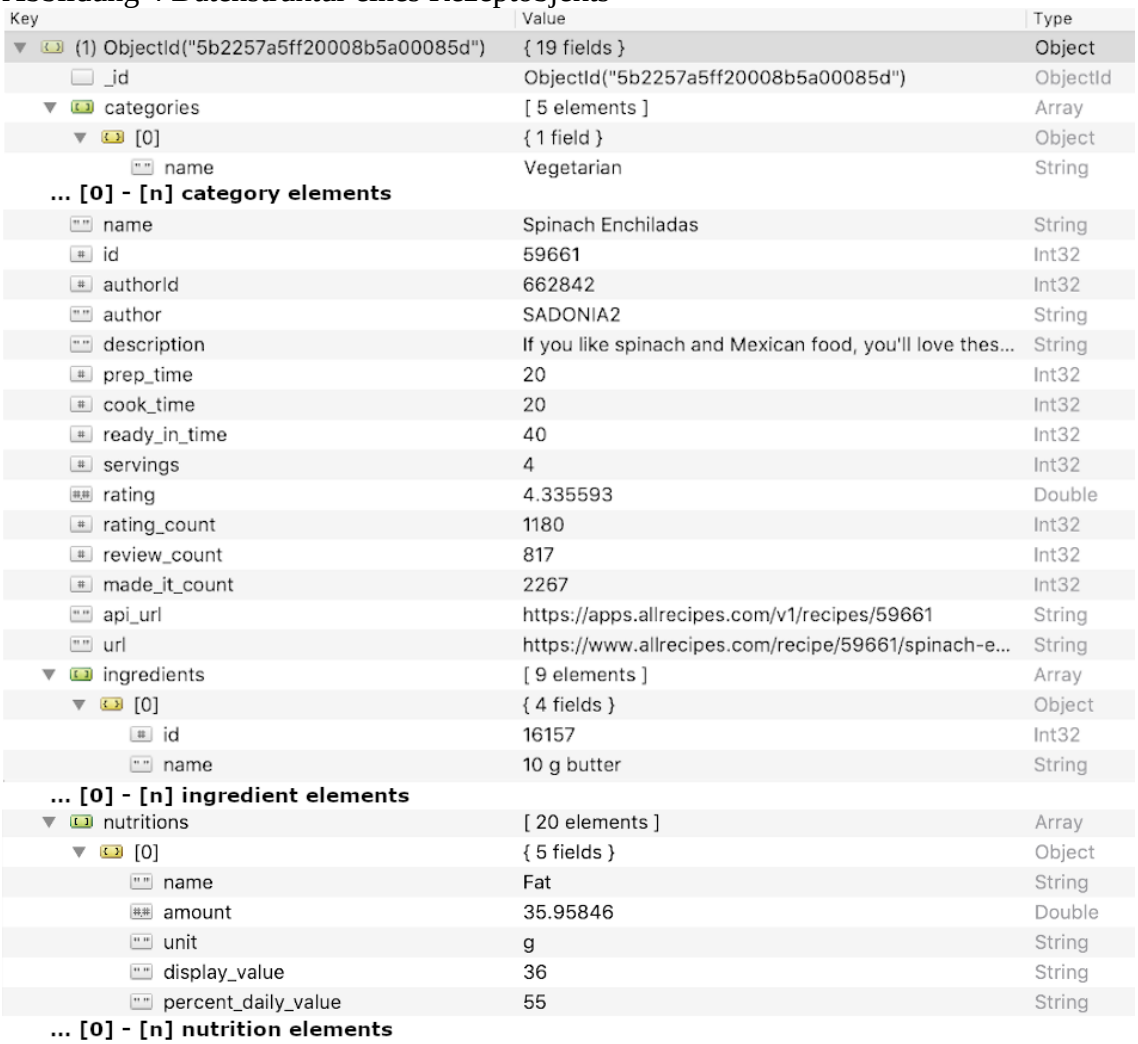
Abbildung 3 Rezeptdarstellung Allrecipe.com



Darstellung der drei Bereiche Zutaten, Anweisungen und Nährstoffe eines Rezepts (Quelle: https://www.allrecipes.com/recipe/17897/hungarian-mushroom-soup, Zugriff: 12.03.2020*)*

Ein Rezeptobjekt hat insgesamt 18 Eigenschaften (keys) denen die Werte (values) der einzelnen Rezepte zugewiesen sind. Es werden aber nicht alle 18 Eigenschaften eines jeden Rezeptes für Berechnungen in dieser Arbeit verwendet. Eigenschaften wie z. B. der Benutzername des Rezepterstellers (*author*) oder die Rezeptbeschreibung (*description*) wurden der Vollständigkeit halber mitgespeichert, da bei der Erstellung des Bot-Programms zum Herunterladen der Rezept-Daten noch nicht endgültig feststand, welche Eigenschaften für die Hypothesen dieser Arbeit von Interesse sind.

Abbildung 4 Datenstruktur eines Rezeptobjekts



Key	Value	Type
(1) ObjectId("5b2257a5ff20008b5a00085d")	{ 19 fields }	Object
_id	ObjectId("5b2257a5ff20008b5a00085d")	ObjectId
categories	[5 elements]	Array
[0]	{ 1 field }	Object
name	Vegetarian	String
... [0] - [n] category elements		
name	Spinach Enchiladas	String
id	59661	Int32
authorId	662842	Int32
author	SADONIA2	String
description	If you like spinach and Mexican food, you'll love thes...	String
prep_time	20	Int32
cook_time	20	Int32
ready_in_time	40	Int32
servings	4	Int32
rating	4.335593	Double
rating_count	1180	Int32
review_count	817	Int32
made_it_count	2267	Int32
api_url	https://apps.allrecipes.com/v1/recipes/59661	String
url	https://www.allrecipes.com/recipe/59661/spinach-e...	String
ingredients	[9 elements]	Array
[0]	{ 4 fields }	Object
id	16157	Int32
name	10 g butter	String
... [0] - [n] ingredient elements		
nutritions	[20 elements]	Array
[0]	{ 5 fields }	Object
name	Fat	String
amount	35.95846	Double
unit	g	String
display_value	36	String
percent_daily_value	55	String
... [0] - [n] nutrition elements		

Datenstruktur eines Rezeptobjekts wie es in der Datenbank gespeichert ist (Quelle: Eigene Darstellung aus dem Programm Robo 3T)

Die drei Eigenschaften ‚Kategorie' (*categories*), ‚Zutaten' (*ingredients*) und ‚Nährwerte' (*nutritions*) sind in einer verschachtelten Struktur abgebildet, da sie mehr als nur einen Wert besitzen. So hat z. B. jedes Nährwert-Objekt (*nutritions*) eines Rezeptes die Eigenschaften ‚Name' (*name*), ‚Menge' (*amount*), ‚Einheit' (*unit*), ‚dargestellter Wert', (*display_value*) und ‚empfohlener Tagesbedarf' (*percent_daily_value*). Zudem hat jedes gespeicherte Rezept Angaben zu den folgenden 20 Nährwertinformationen: Fett, gesättigte Fettsäuren, Eiweiß, Kohlenhydrate, Ballaststoffe, Zucker, Kalorien, Kalorien aus Fett, Cholesterin, Magnesium, Kalzium, Eisen, Vitamin A, Vitamin B1, Vitamin B6, Vitamin C, Nikotinsäure Äquivalente, Natrium, Folsäure und Kalium.

Angaben zu den verwendeten Zutaten sind im Zutaten-Objekt eines Rezepts abgespeichert. Jedes Zutaten-Objekt hat die Eigenschaften ‚Zutaten-ID' (*id*), ‚Zutatenname', (*name*) und ‚Zutatenmenge in Gramm' (*grams*).

4.2 Datenexploration

Die Datenbasis bilden 6305 Rezepte der Webseite Allrecipes.com. Dabei ist jedes dieser Rezepte als ein JSON-Rezeptobjekt abgebildet. In Tabelle 1 sind alle numerischen Eigenschaften (keys) der 6305 Rezepte und ihre statistischen Kennzahlen dargestellt.

Betrachtet man exemplarisch die in Tabelle 1 dargestellte Eigenschaft Kochzeit (*cook_time in min*), so fällt auf, dass die Standardabweichung (std) beinahe doppelt so groß wie der Durchschnittswert (mean) ist. Dies deutet auf eine große Streuung der Werte der Eigenschaft *cook_time in min* hin. Schaut man zudem auf den größten Wert (max), der bei 1500 min liegt, so bestätigt sich diese Annahme. Eine realistischere Einschätzung für die Kochdauer eines durchschnittlichen Rezeptes liefert die Kennzahl Median (median), die hier bei 30 min liegt. Insgesamt lässt sich das für alle numerischen Eigenschaften außer für die 5-Sternebewertung (*rating 1-5*) feststellen.

Des Weiteren fallen die drei Nullwerte für die Kennzahl Minimalwert (min) der Eigenschaften Vorbereitungszeit (*prep_in_time in min*), Kochzeit (*cook_time in min*) und Gesamtzeit (*ready_in_time in min*) auf. Dies deutet darauf hin, dass es zumindest einen oder mehrere Nullwerte in den Rezeptdaten vorhanden sind. Diese Nullstellen sind jedoch nicht unbedingt problematisch, da die Eigenschaften aus Tabelle 1 nicht direkt für die

Berechnungen von Ähnlichkeiten zwischen Rezepten verwendet werden, sondern einen Indikator für die Qualität von Rezepten darstellen. Ein Rezept mit hoher Qualität ist z. B. ein Rezept, das zum einen eine hohe Bewertung (*rating 1-5*) und zum anderen eine hohe Anzahl an Bewertungen (*rating_count*) hat. Ein weiterer Indikator für Qualität ist die Angabe wie oft ein Rezept nachgekocht wurde (*made_it_count*).

Um eine allgemeine Aussage über die durchschnittliche Rezeptqualität der Datenbasis machen zu können, ist es nützlich die Eigenschaften *rating* und *rating_count* in Tabelle 1 zu betrachten. Die Rezepte der Datenbasis haben im Durchschnitt und um Median eine Bewertung von 4,4. Das bedeutet zum einen, dass das durchschnittliche Rezept eine sehr gute Bewertung von 4,4 hat und zum anderen, dass die Bewertungen kaum Ausreißer haben, da die echte Mitte (median) und der Durchschnittswert (mean) identisch sind. Eine gute Bewertung alleine ist aber nicht aussagekräftig genug, da es auch sein kann, dass jedes Rezept z. B. nur einmal gut bewertet worden ist. Deshalb muss man bei der Bewertung der Rezeptqualität auch die Eigenschaft *rating_count* einbeziehen. Die Eigenschaft *rating_count* liegt hier im Durchschnitt bei 253 Bewertungen pro Rezept. Das bedeutet, dass man den Bewertungen der Rezepte durchaus trauen kann, da sie von genug Menschen bewertet worden und nicht zufällig sind.

Tabelle 1 Numerische Eigenschaften der Rezept-Datenbasis

	prep_ time in min	cook_ time in min	ready_in_ time in min	rating 1 – 5	rating_ count	review_ count	made_it_ count
mean	17	59	109	4,4	253	189	372
median	15	30	50	4,4	105	82	143
std	32	105	253	0,3	491	354	761
min	0	0	0	2,0	2	1	5
max	1500	1500	10290	5,0	7898	5588	14356

(Quelle: Eigene Darstellung und Berechnungen, n = 6305 Rezepte)

Tabelle 2 Numerische Eigenschaften der Nährwert-Datenbasis

	Kalorien (kcal)	Fett (g)	Eiweiß (g)	Kohlenhydrate (g)
mean	448,56	23,28	28,05	31,09
median	415,16	20,32	26,47	27,36
std	211,14	15,88	13,70	23,74
min	13,63	0,10	0,16	0,00
max	4709,20	383,92	273,22	236,72

(Quelle: Eigene Darstellung und Berechnung, n = 6305 Rezepte)

Tabelle 2 stellt die numerischen Eigenschaften für die Makronährwerte der Datenbasis dar. Betrachtet man die minimal (min) und maximal (max) Werte sowohl für *Kalorien* als auch für die drei Makronährwerte *Fett*, *Eiweiß* und *Kohlenhydrate* so zeigt sich, dass die Daten Ausreißer aufweisen. Die Durchschnittwerte der drei Makronährwerte sind für den Zweck dieser Arbeit etwas zu niedrig. Zwar sind die Standardabweichungen recht hoch, was eine größere Streuung der Werte bedeutet, jedoch wäre höher Durchschnittwerte von Vorteil. Auf dieses Problem wird im Kapitel Datenbereinigung näher eingegangen.

4.3 Datenimport und Datenstruktur

Für die Manipulation und Analyse der Datenbasis wurde in dieser Arbeit hauptsächlich die Programmbibliotheken pandas, scipy und numpy der Programmiersprache Python 3.7.4 verwendet.

Um eine Manipulation und Analyse der Datenbasis zu ermöglichen, müssen die Daten zunächst aus der MongoDB Datenbank in die Python Programmierumgebung geladen werden. Da die Daten auf der Datenbank im nicht tabellarischen JSON-Format als JSON-Objekte abgespeichert sind, können sie nicht direkt bearbeitet werden. Eine nicht normalisierte Repräsentation der Datenbasis zeigt Tabelle 3. Hier wird jedes Rezept mit seinen Zutaten- und Nährwert-Objekten in einer flachen Darstellung zeilenweise abgebildet. In dieser Form können die Daten nicht sinnvoll bearbeitet und müssen deshalb in eine normalisierte Darstellung gebracht werden.

Tabelle 3 Auszug aus nicht-normalisierter Rezepttabelle

name	id	ingredients	nutritions
Spinach Enchiladas	59661	[{'id': 16157, 'name': '10 g butter', 'grams': 11.36, 'type': 'Normal'}, {'id': 4405, 'name': '40 g sliced green onions', 'grams': 41.8, 'type': '...	[{'name': 'Fat', 'amount': 35.95846, 'unit': 'g', 'display_value': '36', 'percent_daily_value': '55'}, {'name': 'Calories', 'amount': 509.8029, 'u...
Stuffed Peppers	16330	[{'id': 3103, 'name': '305 g ground beef', 'grams': 302.6667, 'type': 'Normal'}, {'id': 1650, 'name': '60 g uncooked long grain white rice', 'gram...	[{'name': 'Fat', 'amount': 9.378057, 'unit': 'g', 'display_value': '9.4', 'percent_daily_value': '14'}, {'name': 'Calories', 'amount': 247.6727, '...
Crispy and Tender Baked Chicken Thighs	235151	[{'id': 10536, 'name': 'cooking spray', 'grams': 0.133, 'type': 'HideAmounts'}, {'id': 6522, 'name': '4 bone-in chicken thighs with skin', 'grams'...	[{'name': 'Fat', 'amount': 11.88371, 'unit': 'g', 'display_value': '11.9', 'percent_daily_value': '18'}, {'name': 'Calories', 'amount': 189.7614, ...
Simple Baked Chicken Breasts	240208	[{'id': 6494, 'name': '4 skinless, boneless chicken breast halves', 'grams': 472.0, 'type': 'Normal'}, {'id': 6307, 'name': '30 ml olive oil', 'gr...	[{'name': 'Fat', 'amount': 9.5803, 'unit': 'g', 'display_value': '9.6', 'percent_daily_value': '15'}, {'name': 'Calories', 'amount': 190.5575, 'un...
Delicious Egg Salad for Sandwiches	147103	[{'id': 16317, 'name': '8 eggs', 'grams': 400.0, 'type': 'Normal'}, {'id': 6294, 'name': '120 ml mayonnaise', 'grams': 110.0, 'type': 'Normal'}, {...	[{'name': 'Fat', 'amount': 31.85464, 'unit': 'g', 'display_value': '31.9', 'percent_daily_value': '49'}, {'name': 'Calories', 'amount': 343.8128, ...

(Quelle: Eigene Darstellung)

Um die Daten in ein geeignetes tabellarisches Format zu überführen, müssen die Daten mit Hilfe einer Normalisierung in mehrere Tabellen aufgespalten werden. Für den Verwendungszweck dieser Arbeit wurde die Datenbasis auf die drei Tabellen *ingredients*, *nutritions* und *recipe_db* aufgeteilt. Wobei der Primärschlüssel jeder Tabelle die Rezept ID eines Rezeptes ist.

Tabelle 4 Auszug aus der Tabelle *ingredients*

id	ingredients_id	ingredients_name	ingredients_grams
59661	16157	10 g butter	11,36
59661	4405	40 g sliced green onions	41,80
59661	4342	1-1/2 cloves garlic, minced	4,80
59661	4520	3/4 (10 ounce) package frozen chopped spinach	227,20
59661	16243	180 g ricotta cheese	182,40
59661	16261	90 g sour cream	92,00

(Quelle: Eigene Darstellung)

Die Tabelle *ingredients* beinhaltet alle Informationen des Zutaten-Objekts eines Rezeptes. Das *ingredients* Zutaten-Objekt hat die Spalten *id* (Rezept-ID), *ingredients_id*, *ingredients_name* und *ingredients_grams* wie in Tabelle 4 dargestellt.

Tabelle 5 Auszug aus der Tabelle *nutritions*

name id	Calcium	Calories	Calories from Fat	Carbohyd-rates	Cho-lesterol	Dietary Fiber	Fat
6900	117,80	105,27	21,33	17,99	1,63	0,57	2,37
7198	244,17	731,51	325,65	73,30	88,69	4,64	36,18
8493	415,52	454,28	177,75	23,83	203,87	2,00	19,75
8494	58,09	834,44	514,68	4,77	283,75	0,64	57,19

(Quelle: Eigene Darstellung)

Die Tabelle *nutritions* folgt derselben Logik und beinhaltet alle Informationen des Nährwert-Objekts für jedes Rezept. Die Spalten der Tabelle *nutritions* sind *id* (Rezept-ID) und 20 Nährwertinformationen: Fett, gesättigte Fettsäuren, Eiweiß, Kohlenhydrate, Ballaststoffe, Zucker, Kalorien, Kalorien aus Fett, Cholesterin, Magnesium, Kalzium, Eisen,

Vitamin A, Vitamin B1, Vitamin B6, Vitamin C, Nikotinsäure Äquivalente, Natrium, Folsäure und Kalium.

Die dritte Tabelle *recipe_db* ist primär für die Berechnung der Jaccard Distanz notwendig. Da die Jaccard Distanz – generell gesagt – Aussagen über das Vorhandensein bzw. das Fehlen von Zutaten in Rezepten trifft, erwartet die Methode zur Berechnung der Jaccard Distanz eine Datenstruktur, die binär aufgebaut ist.

Tabelle 6 Auszug aus der Tabelle *recipes_db*

id	111	126	257	443	445	578	615	629	631	858	...	23047	23274
7198	0	0	0	0	0	0	0	0	0	0	...	0	0
8493	0	0	0	0	0	0	0	0	0	0	...	0	0
8494	0	0	0	0	0	0	0	0	0	0	...	0	0
8495	0	0	0	0	0	0	0	0	0	0	...	0	0
8496	0	0	0	0	0	0	0	0	0	0	...	0	0
8497	0	0	0	0	0	0	0	0	0	0	...	0	0
8498	0	0	0	0	0	0	0	0	0	0	...	0	0
8500	0	0	0	0	0	0	0	0	0	0	...	0	0
8503	0	0	0	0	0	0	0	0	0	0	...	0	0
8506	0	0	0	0	0	0	0	0	0	0	...	0	0

(Quelle: Eigene Darstellung)

Für die Erstellung der Tabelle *recipes_db* sind die Informationen der Tabelle *ingredients* notwendig, da sie Informationen über die Zusammensetzung eines jeden Rezeptes hinsichtlich seiner Zutaten besitzt.

Mit den Informationen der *ingredients* Tabelle wurde die Tabelle *recipes_db* generiert. Die Tabelle *recipes_db* ist eine Matrix, deren Spalten aus Zutaten IDs und ihre Zeilen aus Rezept IDs bestehen. Die Zusammensetzung von allen Datenbank-Rezepten ist in dieser Tabelle in binärer Form abgespeichert. Das bedeutet, wenn man z. B. nachvollziehen möchte welche Zutaten in einem Rezept vorhanden sind, kann man in einer Zeile die Spalten suchen, die eine ‚1' haben und somit die Spalten der Zutaten IDs identifizieren.

4.4 Datenbereinigung

Im ersten Schritt der Datenbereinigung werden zunächst alle ungültigen Werte (NA) in allen Tabellen entfernt. Zudem müssen alle Rezepte, die die Zutaten ID ‚0' aufweisen gefiltert und aus der Rezept-Datenbank als auch aus der Nährwert-Datenbank entfernt werden. Das ist vor allem deswegen wichtig, weil Rezepte oft nicht nur eine fehlerhafte Zutat mit ID ‚0' hatten, sondern mehrere gleichzeitig. Nach der Bereinigung der Rezepte mit den fehlerhaften IDs hat sich die Gesamtzahl der Rezepte in der Rezept-Datenbank auf von 6305 auf 6029 Rezepte reduziert.

Der zweite Schritt der Bereinigung hat den größten Einfluss auf die Rezeptanzahl. In diesem Schritt werden die Zutaten hinsichtlich ihrer Mengen betrachtet. Zutaten, die sehr selten vorkommen (<=5) können sich negativ auf die Berechnung der Rezept-Empfehlung auswirken und werden deshalb entfernt. Es werden jedoch nicht nur die Zutaten entfernt, sondern auch alle Rezepte samt ihrer Nährwertinformation. Nach dieser Bereinigung hat sich die Anzahl der Rezepte auf 4062 verkleinert.

Tabelle 7 Top 10 Liste der Standardzutaten

ingredients_id	ingredients_name	ing_count
16421	salt and pepper to taste	2126
4342	1-1/2 cloves garlic, minced	1506
4397	1/2 extra large onion, chopped	1412
16406	0.3 g ground black pepper	1185
16157	10 g butter	1016
6307	30 ml olive oil	944
6494	4 skinless, boneless chicken breast halves	786
2496	160 ml water	749
16238	80 g grated Parmesan cheese	574
16317	8 eggs	538

(Quelle: Eigene Berechnung)

Im dritten Schritt werden alle Zutaten entfernt, die zu häufig in den Rezepten vorkommen. Diese Zutaten sind sogenannte Standardzutaten, die oft zu einem Rezept

dazugehören wie z. B. Salz, Pfeffer und Wasser. Tabelle 7 zeigt die zehn häufigsten Zutaten und wie häufig sie vorkommen. Da die Standardzutaten sehr häufig in Rezepten vorkommen, erhöhen sie zwar die Ähnlichkeit zwischen Rezepten, dies jedoch auf einer Ebene, die nicht den Kern eines Rezeptes ausmacht. Da bei der Berechnung von Rezepten alle Zutaten gleich gewichtet sind, ist es wichtig Standartzutaten aus den Rezepten zu filtern, um den Ähnlichkeitswert zwischen zwei Rezepten nicht zu entkräften. Es werden jedoch nur die Zutaten aus den Rezepten entfernt und nicht die Rezepte selbst, da sich sonst die Anzahl der Rezepte zu stark minimieren würde.

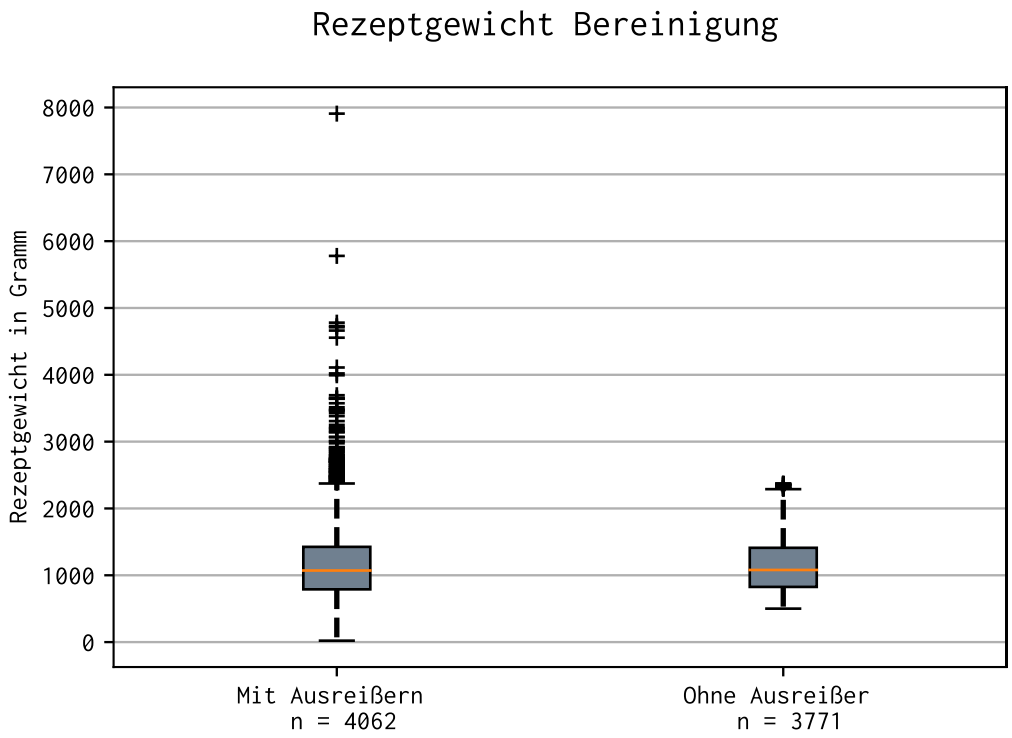
Die letzte Datenbereinigung betrifft die Nährwerte der Rezept-Datenbank. Die Bereinigung findet jedoch nicht direkt über die Nährwerte statt, sondern über einen Umweg mit Hilfe des Gewichts eines jeden Rezeptes. Wieso diese Bereinigung notwendig wurde, erklärt der nächste Absatz.

Die anfängliche Idee bestand darin die Vergleichbarkeit von Nährwertabgaben eines jeden Rezeptes so zu garantieren, dass man jedes Rezept auf einen Basisgramm Wert seiner Zutaten zurückrechnet und die jeweiligen Nährwertabgaben prozentuell an diese Basisgramm Angabe anpasst. Das Ziel war es die Gramm Mengen der Zutaten und ihren entsprechenden Nährwerten vergleichbar mit dem low-fat und dem low-carb Profil zu machen, da die beiden Profile sich an den Nährwertangaben einer Hauptmahlzeit orientieren. Dieses Ziel konnte aufgrund eines Fehlers beim Herunterladen der Rezept-Daten von der Allreccipes.com Webseite leider nicht erreicht werden. Für das Runterladen der Daten, wurden das Bot-Programm so programmiert, dass die Portionsgröße für alle Rezepte auf 4 festgelegt ist. Der Hintergedanke dabei war, dass eine gleiche Portionsgröße für alle Rezepte eine gewisse Vergleichbarkeit garantiert. Es wurde angenommen, dass die Nährwertangaben sich an der eingestellten Portionsgröße automatisch neu berechnen. Dies war leider nicht der Fall und führte dazu, dass die Nährwertangaben eines Rezeptes sich an der Portionsgröße bemessen, die der ursprüngliche Autor anfänglich festgelegt hat. Da beim Runterladen der Daten die ursprüngliche Portionsgröße nicht mitgespeichert, sondern auf 4 umgestellt wurde, fehlt die Verbindung zwischen den Nährwerten und der Grammanzahl der Rezepte. Somit ist eine Normalisierung der Nährstoffe auf eine Basis

Grammanzahl einer Hauptmahlzeit nicht möglich. Die vorhandenen Nährwert-Daten sind immer als Angaben pro eine Portion abgespeichert. Bei mehreren Portionen bleibt es allerdings schwer bzw. nicht nachvollziehbar.

Da das besprochene Problem im Kern nicht gelöst werden kann, wurde eine Herangehensweise gewählt, die es zumindest abmildern soll.

Abbildung 5 Box-Plot-Diagramm Rezeptgewicht



(Quelle: Eigene Darstellung)

Um die Rezepte herauszufiltern, deren Nährwerte sich als Ausreißer darstellen, wurde das Gewicht für jedes Rezept über die Grammangaben der Zutaten berechnet.

Abbildung 5 zeigt zwei Box-Plots in einem Diagramm. Der linke Box-Plot zeigt die Verteilung des Rezeptgewichts mit den Ausreißern, die als Plus-Symbol dargestellt sind. Ein Rezept mit dem Gewicht von beinahe acht Kilogramm ist der größte Ausreißer. Das Kriterium für die Definition eines Ausreißers ist die gebräuchliche Whisker-Länge

von 1,5*IQR (*interquartile range oder Interquartilsabstand*). Der Interquartilsabstand ist im Diagramm als graue Box dargestellt. In dieser Box liegen die mittleren 50 % aller Rezepte. Auf der rechten Seite des Diagramms ist der Box-Plot nach der Bereinigung abgebildet. Alle Rezepte, deren Gewicht größer 1,5*IQR war, wurden aus der Datenbank entfernt. Zusätzlich wurde das Anfangsgewicht von 0 Gramm auf 500 Gramm gesetzt. Nach dieser Bereinigung beträgt die Gesamtzahl der Rezepte 3771.

5 Modellbildung

Dieses Kapitel beschreibt das Modell, das verwendet wurde, um die Ähnlichkeiten zwischen den Rezepten zu berechnen. Es handelt sich um ein zweistufiges Modell. Im ersten Schritt wird die Zutaten-Ähnlichkeit berechnet. Im zweiten Schritt die Nährwert-Ähnlichkeit von Rezepten bezogen auf ein Nährwert-Profil. Die in dieser Arbeit benutzten Nährwert-Profile sind zum einen ein low-fat-Profil (wenig Fett) und zum anderen ein low-carb-Profil (wenige Kohlenhydrate).

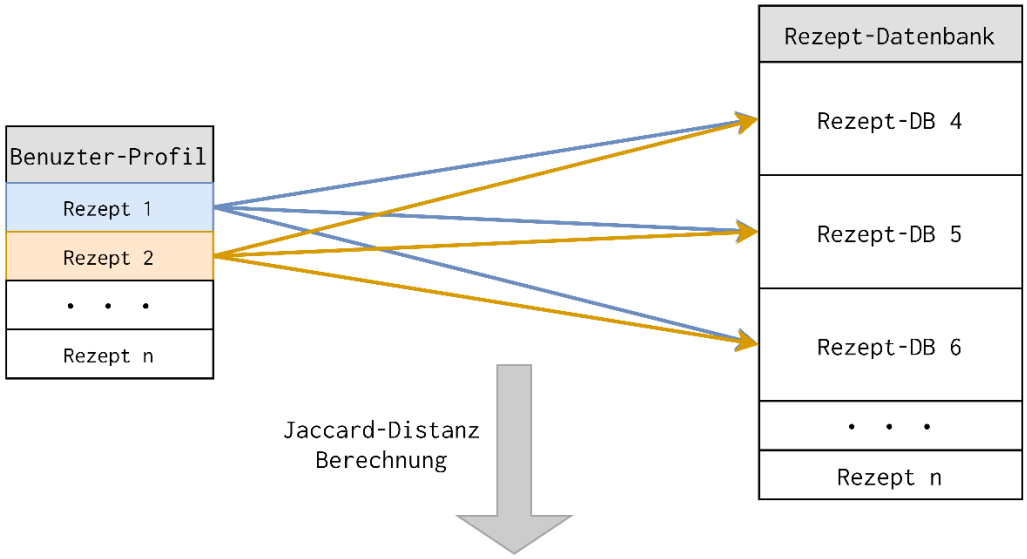
5.1 Hybride Methode

Die in dieser Arbeit entwickelte Methode der Hybriden Ähnlichkeit ist ein zweistufiges Verfahren bei der zunächst die Zutaten-Ähnlichkeit zwischen einem Benutzer-Profil und allen Rezepten aus der Rezept-Datenbank berechnet wird. Im zweiten Schritt werden diese Ergebnisse genutzt, um eine Nährwert-Ähnlichkeit zu bestimmen.

5.2 Zutaten-Ähnlichkeit

Bei der Berechnung der Zutaten-Ähnlichkeit wird die Jaccard Distanz von jedem einzelnen Rezept des Benutzerprofils zu allen Rezepten der Rezept-Datenbank ermittelt. Dies ist in Abbildung 6 dargestellt. Das Ergebnis ist eine N x M Distanz-Matrix, wobei N die Anzahl der Rezepte im User-Profil und M die Anzahl der Rezepte in der Rezept-Datenbank darstellt. Die errechnete Distanz zwischen zwei Rezepten kann einen Wert zwischen 0 und 1 annehmen.

Abbildung 6: Berechnung der Jaccard Distanz-Matrix



Beispiel-Tabelle: Jaccard-Distanz-Matrix

	Rezept-DB 4	Rezept-DB 5	Rezept-DB 6
Rezept 1	0	1	0,5
Rezept 2	1	0	0,7

(Quelle: Eigene Darstellung)

Dieser Wert ist die Jaccard Distanz, die den Abstand zwischen zwei Rezepten angibt. Der Abstand wird anhand der in beiden Rezepten vorhandenen Zutaten berechnet. Wobei 0 eine perfekte Überschneidung der verwendeten Zutaten in beiden Rezepten darstellt und eine 1 ein komplett unterschiedliches Rezept beschreibt, das keine einzige Zutat gemein hat. Die Top 10 Rezeptliste ist absteigend sortiert, so dass das erste Rezept die kürzeste Distanz und somit die größte Ähnlichkeit besitzt. Da es unterschiedliche Möglichkeiten gibt mithilfe der Jaccard Distanz die Distanz bzw. Ähnlichkeit zwischen zwei Rezepten zu berechnen, werden im nächsten Abschnitt drei unterschiedliche Methoden vorgestellt, um die Jaccard Distanz zu berechnen. Die Ergebnisse dieser Berechnungen werden miteinander und mit Zufallswerten verglichen und im Ergebnis-Kapitel dieser Arbeit diskutiert.

5.2.1 Naive Berechnung

Ausgangspunkt der Naiven Berechnung ist eine Jaccard Distanz-Matrix mit den errechneten Abständen zwischen den Rezepten des Benutzer-Profils und der Rezept-Datenbank. Die Distanzen der Datenbank-Rezepte werden, wie in Abbildung 7 dargestellt, zunächst spaltenweise aufsummiert. Im nächsten Schritt werden alle Summen durch die Anzahl der Rezepte des Benutzer-Profils dividiert, um auf eine Jaccard Distanz zwischen 0 und 1 zu kommen. Im letzten Schritt werden die Distanzen aufsteigend sortiert. Die ersten 10 Rezepte bilden somit die Top 10 Rezepte, die am ähnlichsten zum Benutzer-Profil sind.

Abbildung 7 Naive Berechnung Jaccard Distanz

Beispiel-Tabelle: Jaccard-Distanz-Matrix

	Rezept-DB 4	Rezept-DB 5	Rezept-DB 6
Rezept 1	1	0,6	1
Rezept 2	1	1	0,7
Rezept 3	0,5	0,7	1

1. Spaltenweise aufsummieren
2. Spaltensumme durch Anzahl User-Rezepte (n=3) dividieren

Beispiel-Tabelle: Jaccard-Distanz aufsummiert und durch 3 dividiert

	Rezept-DB 4	Rezept-DB 5	Rezept-DB 6
Jaccard-Distanz	0,83	0,76	0,9

3. Tabelle transponieren
4. Aufsteigend nach Jaccard-Distanz sortieren

Beispiel-Tabelle: Naive Berechnung TOP 3

	Jaccard-Distanz
Rezept-DB 5	0,76
Rezept-DB 4	0,83
Rezept-DB 6	0,9

(Quelle: Eigene Darstellung und Berechnung)

5.2.2 Kürzeste Distanz Berechnung

Ausgangspunkt der Kürzeste Distanz Berechnung ist eine Jaccard-Distanz-Matrix mit den errechneten Abständen zwischen den Rezepten des Benutzer-Profils und der Rezept-Datenbank. Anders als bei der Naiven Berechnung werden die Distanzen jedoch nicht aufsummiert und dividiert, sondern es wird nur ein Datenbank-Rezept – nämlich jenes mit der kürzesten Distanz – pro Spalte ausgewählt. Abbildung 8 veranschaulicht das Verfahren. Das Ergebnis der Berechnung ist eine Top 10 Rezeptliste, die aufsteigend nach der Jaccard Distanz sortiert ist.

Abbildung 8 Kürzeste Distanz Berechnung Jaccard Distanz

Beispiel-Tabelle: Jaccard-Distanz-Matrix

	Rezept-DB 4	Rezept-DB 5	Rezept-DB 6
Rezept 1	1	0,6	1
Rezept 2	1	1	0,7
Rezept 3	0,5	0,7	1

1. Datenbank-Rezept mit der kürzesten Distanz auswählen
2. Aufsteigend nach Jaccard-Distanz sortieren

Beispiel-Tabelle: Kürzeste Distanz Berechnung TOP 3

	Jaccard-Distanz
Rezept-DB 4	0,5
Rezept-DB 5	0,6
Rezept-DB 6	0,7

(Quelle: Eigene Darstellung)

5.2.3 Rezept-Vektor Berechnung

Bei der Rezept Vektor Berechnung werden die Jaccard Distanzen zwischen Datenbank-Rezepten und dem Benutzer-Rezept nicht wie bei der Naiven und der Kürzesten-Distanz-Berechnung einzeln pro Benutzer-Rezept berechnet. Stattdessen werden alle Zutaten der Benutzer-Rezepte in einen einzelnen Rezept Vektor übertragen. Der

Benutzer hat demnach nur ein einziges Rezept in seinem Profil, das aber alle Zutaten der zuvor vereinzelten Rezepte beinhaltet. Mit dem einzelnen Rezept Vektor werden die Jaccard Distanzen zu allen Datenbank-Rezepten berechnet. Auch bei dieser Methode ist das Ergebnis eine nach aufsteigender Jaccard Distanz sortierte Top 10 Rezeptliste.

Abbildung 9 Rezept Vektor Bildung

Beispiel-Tabelle: Binäre Darstellung von Benutzer-Rezepten

	Zutaten ID 1	Zutaten ID 2	Zutaten ID 3	Zutaten ID 4
Rezept 1	1	0	1	0
Rezept 2	1	1	0	0
Rezept 3	1	1	0	0

Alle auftretenden Zutaten (Zutat = 1) in ein Rezept zusammenfassen

Beispiel-Tabelle: Binäre Darstellung von Rezept-Vektor

	Zutaten ID 1	Zutaten ID 2	Zutaten ID 3	Zutaten ID 4
Rezept-Vektor	1	1	1	0

(Quelle: Eigene Darstellung)

5.3 Nährwert-Ähnlichkeit

Die Berechnung der Nährwert-Ähnlichkeit ist der zweite Schritt der hybriden Methode und basiert auf den zuvor berechneten Jaccard Distanzen zwischen dem Benutzer-Profil und der Rezept-Datenbank. Die Nährwert-Ähnlichkeit wird jedoch nicht für alle drei Methoden (Naive-, Kürzeste Distanz- und Rezept Vektor Berechnung) durchgeführt, sondern nur für jene mit der besten Performanz. Die Performanz der drei genannten Methoden für die Berechnung der Zutaten-Ähnlichkeit wird im Ergebnisteil bestimmt.

Ausgangspunkt für die Berechnung der Nährwert-Ähnlichkeit ist eine Top 500 Rezeptliste. Die Rezeptliste enthält die 500 ähnlichsten Rezepte zum Benutzerprofil und bildet deshalb die Basis für die Ermittlung der Nährwerte-Ähnlichkeit. Die Anzahl der

Rezepte ist auf 500 festgelegt, da man eine Varianz in den 20 Nährwerten eines Rezeptes benötigt, um eine Nährwert-Ähnlichkeit zum low-fat bzw. low-carb Profil berechnen zu können. Je mehr Rezepte man betrachtet, desto höher ist die Wahrscheinlichkeit, dass man die Nährwertmengen in einem Rezept findet, die man mit den beiden Nährwert-Profilen definiert hat. Gleichzeitig sollte die Größe der Rezeptliste nicht zu groß sein, da damit auch die Wahrscheinlichkeit steigt, dass man eine gute Nährwert-Ähnlichkeit zu einem Rezept hat, das aber keine gute Zutaten-Ähnlichkeit zum Benutzer-Profil besitzt. So könnte z. B. ein Rezept, das auf dem letzten Platz der Rezeptliste ist – somit eine eher schlechte Zutaten-Ähnlichkeit zum Benutzerprofil aufweist – eine gute Nährwert-Ähnlichkeit besitzt oder umgekehrt. Die Größe der Rezeptliste muss als Kompromiss zwischen Zutaten-Ähnlichkeit und Nährwert-Ähnlichkeit verstanden werden.

Tabelle 8 Nährwerte des low-fat und des low-carb Profils

	Kalorien	Fett	Eiweiß	Kohlenhyd-rate	Ballast-stoffe	Eisen	Kalzium
low-fat	716 kcal	10 g	51,75 g	103,7 g	12,67 g	2 mg	266.67 mg
low-carb	716 kcal	41,3 g	53,3 g	10 g	12,67 g	2 mg	266.67 mg

	Magnesium	Vit. A	Vit. B1	Vit. B6	Vit. C	Folsäure	Nikotin-säure
low-fat	116,6 mg	208 µg	0,3 mg	0,36 mg	25 mg	133,3 µg	4 mg
low-carb	116,6 mg	208 µg	0,3 mg	0,36 mg	25 mg	133,3 µg	4 mg

(Quelle: vgl. NCBI, 2020 & vgl. Schnur, 2013, S.39)

Die 20 Nährwerte, die als Referenz für die low-fat und low-carb Profile dienen, basieren auf dem empfohlenen Tagesbedarf an Nährstoffen für Männer zwischen 31 und 50 Jahren des *National Center for Biotechnology Information* (NCBI) (vgl. NCBI, 2020). Die Tagesbedarfsmenge wurde auf drei aufgeteilt, um die Menge für eine Hauptmahlzeit zu erhalten.

Da das NCBI nicht alle notwendigen Nährwertinformationen aufweisen konnte, wurden zudem noch Daten der *Deutschen Gesellschaft für Ernährung e.V.* (DGE) verwendet. Diese Daten beziehen sich auf den empfohlenen Tagesbedarf an Nährstoffen für eine Mittagsverpflegung für Personen zwischen 25 und 51 Jahren (vgl. Schnur, 2013,

S.39). Für die Nährwerte und Nährwertinformationen Natrium, Kalium, Zucker, gesättigte Fettsäuren, Cholesterin und Folsäure sind leider keine Nährwertangeben in den benutzten Quellen vorhanden. Diese Informationen werden deshalb nicht in die Berechnung mit einbezogen.

Die Definitionen des low-fat und low-carb Profils basieren auf einem einfachen Prinzip. Der jeweilige Makronährwert wird auf ein relativ niedriges Niveau von 10 mg festgelegt. Die dadurch frei gewordenen Kalorien werden auf die jeweils anderen beiden Makronährwerte verteilt. So wird z. B. beim low-fat Profil 14g Fett reduziert. Das entspricht 126 kcal (1g Fett = 9 kcal). Die 126 kcal werden im gleichen Teil auf Eiweiß und Kohlenhydrate verteilt, um die Gesamtanzahl der Kalorien pro Mahlzeit gleich zu halten. Die gleiche Rechnung wird auch beim low-carb Profil durchgeführt. Das Ergebnis dieser Rechnungen sieht man im hervorgehobenen Kasten der Tabelle 8. Die restlichen Nährwerte sind für beide Profile identisch.

Um die Nährwert-Ähnlichkeit zwischen den Profilen low-fat und low-carb und den Nährwerten der Rezept-Datenbank zu bestimmen, wird die Euklidische Distanz verwendet. Die Berechnung verläuft nach einem ähnlichen Schema wie die Berechnung der Rezept Vektor Methode. Auch hier wird zu einem einzelnen Vektor (z. B. low-fat Profil) der 14 Nährwerte (siehe Tabelle 8) enthält, die Distanz zu den Nährwerten der Rezept-Datenbank berechnet. Ein Unterschied zur Rezept Vektor Methode ist, dass die jeweiligen Makronährwerte Fett und Kohlenhydrate für die Distanzberechnung gewichtet werden. Das bedeutet, dass diese Werte für die Distanz Berechnung eine Priorisierung bekommen. Es wird also immer zuerst geschaut, dass die im low-fat und low-carb festgelegten Fett- und Kohlenhydratwerte erreicht werden bevor die Berechnung der restlichen Nährwerte stattfindet.

Das Ergebnis ist eine Liste von 10 Rezepten und deren jeweilige Euklidische Distanz. An der ersten Stelle dieser Liste befindet sich schließlich das Rezept, das die ähnlichsten Nährwerte zum low-fat bzw. low-carb Profil und somit die niedrigste Euklidische Distanz besitzt.

6 Vorstellung der Ergebnisse

In diesem Kapitel werden die Ergebnisse dieser Arbeit vorgestellt und diskutiert. Die Zutaten-Ähnlichkeiten werden als erstes für die beiden Benutzer-Profile ‚Fett-Profil' und ‚Normal-Profil' berechnet. Im zweiten Schritt werden mit Hilfe der Nährwert-Ähnlichkeit Rezeptvorschläge mit einem low-fat und einem low-carb Profil generiert. Für das ‚Fett' Benutzer-Profil wird das low-fat Profil gewählt und für das ‚Normal' Benutzer-Profil wird das low-carb Profil verwendet. Die beiden Benutzer-Profile ‚Fett' und ‚Normal' werden im Verlauf dieses Kapitels genauer beschrieben.

Bevor die Ergebnisse dargestellt werden, muss zunächst der Modus der Berechnung definiert und die Kriterien eingeführt werden, mit denen die Performanz der Berechnungen gemessen wird.

Wie in der Einleitung beschrieben ist das Ziel dieser Arbeit die drei folgenden Hypothesen zu überprüfen:

Hypothese 1: *„Wenn ein Rezeptvorschlag mit der Jaccard Distanz generiert wird, dann hat er eine höhere Überschneidung an Zutaten zum Benutzter-Profil als ein zufällig generierter Rezeptvorschlag."*

Hypothese 2: *„Wenn ein Rezeptvorschlag mit der Euklidischen Distanz berechnet wird, dann sind die Nährwerte der Rezepte näher an den Nährwerten des low-fat bzw. low-carb Profils als ein zufällig generierter Rezeptvorschlag."*

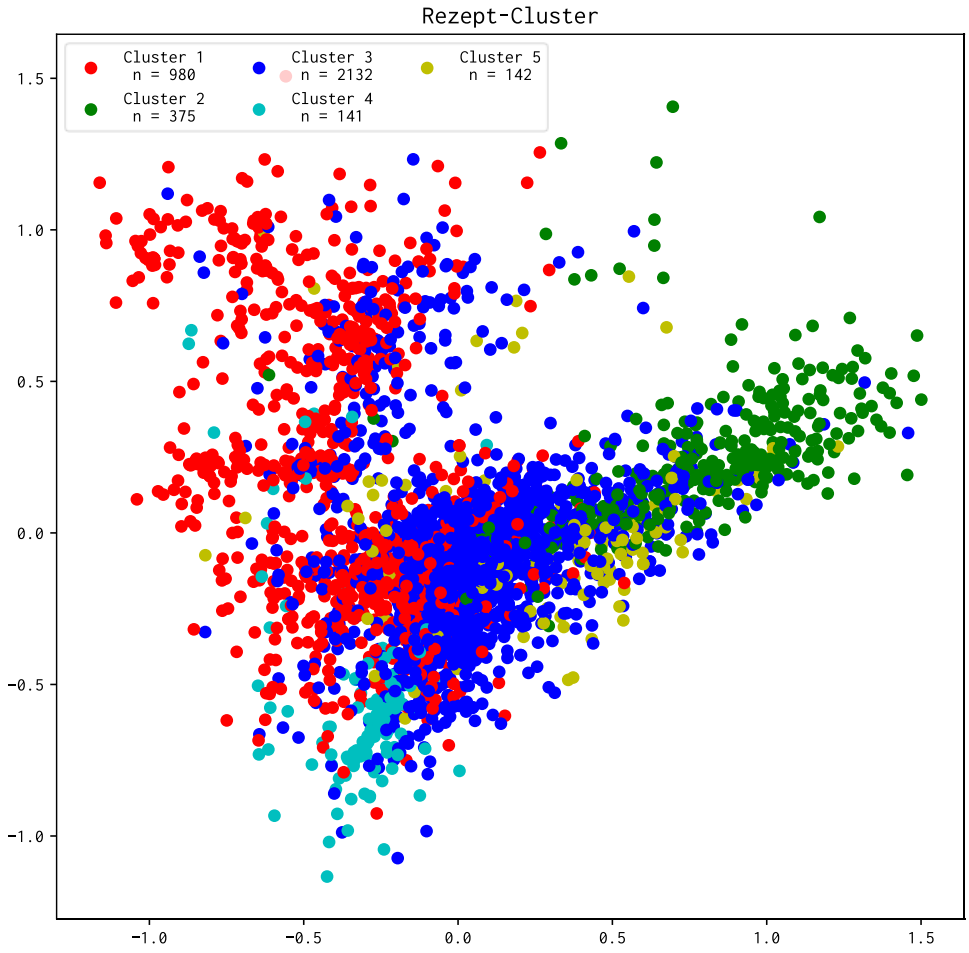
Hypothese 3: *„Je mehr Rezepte ein Benutzer in seinem Profil besitzt, desto höher ist die Überschneidung zwischen den Benutzter-Profil-Zutaten und den Rezeptvorschlags-Zutaten."*

Zur Überprüfung der ersten Hypothese wurden die drei Methoden Naive-, Kürzeste Distanz- und Rezept Vektor-Berechnung herangezogen. Die drei Methoden werden

an den Ergebnissen einer Zufalls-Berechnung gemessen, um Unterschiede und Performanz zu vergleichen. Die zweite Hypothese wird mit den Ergebnissen aus der Zutaten-Ähnlichkeits-Berechnung überprüft. Die dritte Hypothese wird anhand der Distanzberechnungen zu den Profilen low-fat und low-carb getestet.

6.1 Aufbau der Berechnungen

Abbildung 10 Darstellung der fünf Rezept-Cluster



(Quelle: Eigene Darstellung)

Für das Überprüfen der drei Hypothesen mussten für das Fett-Profil und das Normal-Profil Rezepte ausgewählt werden. Da die Rezept-Daten keine Informationen zu echten Benutzer-Profilen und ihren Rezept-Kollektionen besitzen, wurden für die Auswahl

der Rezepte nicht einfach zufällige Rezepte aus der Rezept-Datenbank gewählt, sondern folgende Methode benutzt. Wie in Abbildung 10 dargestellt, wurden mit Hilfe des agglomerativen Clusterverfahrens fünf Cluster basierend auf der Rezept-Ähnlichkeit gebildet. Die Überlegung dahinter bestand darin, dass echte Benutzer nicht einfach zufällige Rezepte in ihrer Kollektion besitzen, sondern Rezepte, die eine gewisse Ähnlichkeit aufweisen.

Aus den fünf erstellen Rezept-Clustern wurde für das Fett-Profil das Cluster 1 mit 980 Rezepten gewählt. Für das Normal-Profil wurde das Cluster 2 mit 375 Rezepten bestimmt. Das größere Cluster 1 wurde deshalb für das Fett-Profil gewählt, da die Rezepte im Fett-Profil eine spezielle Bedingung erfüllen müssen: Die Rezepte des Fett-Profils müssen einen Fettgehalt zwischen 35 und 40 Gramm pro Rezept aufweisen, was überdurchschnittlich hoch ist und deshalb eine größere Auswahl an Rezepten verlangt. Das Normal-Profil unterliegt keinen speziellen Beschränkungen, da es das Profil eines durchschnittlichen Benutzers abbilden soll. Für beide Profile wurde die Rezeptanzahl auf maximal 40 Rezepte begrenzt.

Sowohl für das Fett- als auch für das Normal-Profil wurde die Zutaten-Ähnlichkeit mit den Methoden Naive-, Kürzeste-Distanz- und Rezept Vektor Berechnung ermittelt. Die Berechnungen werden für beide Profile mit unterschiedlicher Anzahl (5, 10, 20 und 40 Rezepte) an Benutzer-Profil-Rezepten durchgeführt. Diese Informationen helfen dabei die dritte Hypothese zu testen.

Um die Reliabilität der ermittelten Performanz-Kriterien zu erhöhen, wurde jede Berechnung drei Mal durchgeführt und ein Durchschnitt aus den Ergebnissen gebildet. Zudem wurde für die drei Wiederholungen der Messung immer das gleiche Sample von 40 Rezepten verwendet. Die Ziehung der 5, 10 und 20 Rezepte aus dem Sample mit 40 Rezepten war jedoch immer zufällig.

Für die Erstellung der Zufalls-Berechnung wurde dasselbe Rezept-Sample verwendet wie für die drei Methoden der Zutaten-Ähnlichkeit. Die Top 10 Liste der Rezept-Empfehlungen wurde dementsprechend zufällig generiert.

6.2 Performanz-Kriterien

Da diese Arbeit keine empirischen Informationen über die Güte der Rezeptvorschläge besitzt, müssen andere Kriterien für die Bewertung herangezogen werden. Für diesen Zweck wurden folgende Kriterien definiert:

1. *Überschneidungszutaten:*
   Die Überschneidungszutaten sind die Zutaten, die sowohl im Benutzer-Profil als auch in der Top 10 Rezeptvorschlags-Liste vorkommen. Zutaten, die sich mehrfach überschneiden, werden nur einmal gezählt.

2. *Aufsummierte Überschneidungszutaten:*
   Hier werden alle Überschneidungszutaten aufsummiert, auch die Zutaten, die sich in mehreren Rezepten überschneiden.

3. *Einzigartige Zutaten:*
   Bei diesem Kriterium wird die Anzahl der einzigartigen Zutaten in der Top 10 Rezeptvorschlags-Liste angegeben.

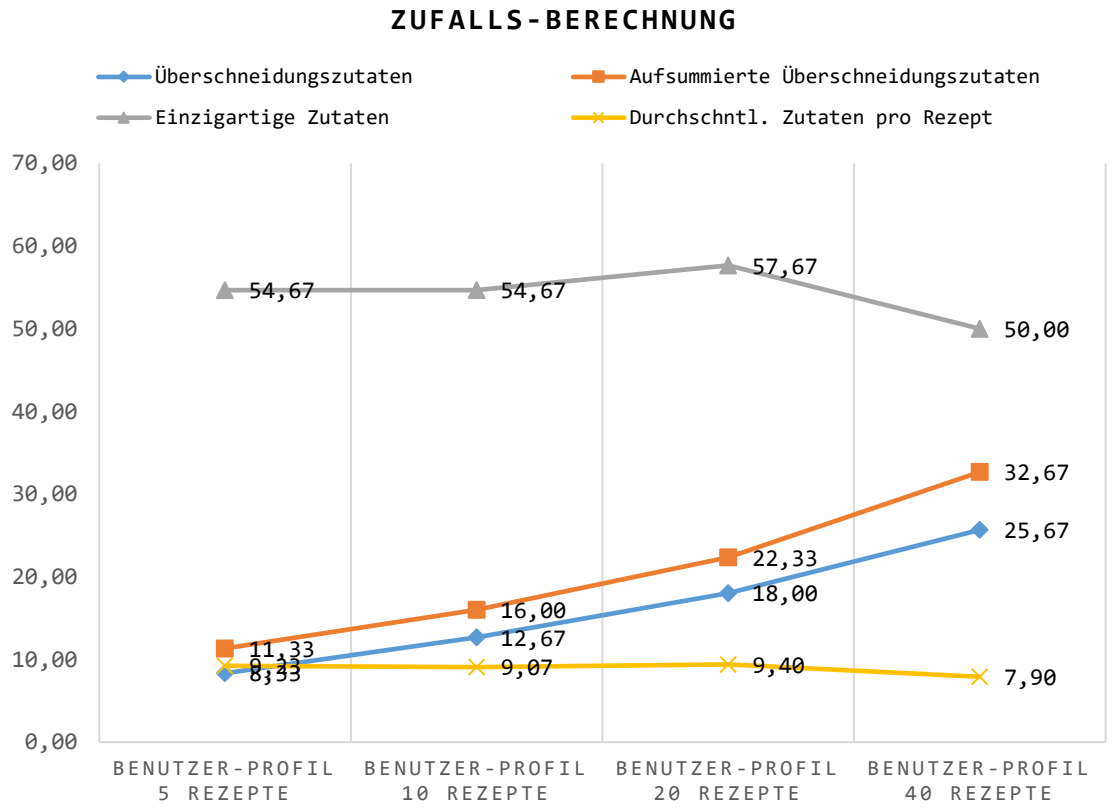
4. *Durchschnittliche Anzahl an Zutaten pro Rezept:*
   Dieses Kriterium gibt die durchschnittliche Anzahl an Zutaten der Top 10 Rezeptvorschlags-Liste an. Es ist speziell für die Rezept Vektor Methode notwendig.

6.3 Ergebnisse Zutaten-Ähnlichkeit

Im Folgenden wird nur die Zutaten-Ähnlichkeit für das Fett-Profil besprochen, da sich die Berechnungen für das Normal-Profil nicht signifikant unterscheiden. Die Ergebnisse für das Normal-Profil werden dennoch benötigt, da sie später für die Berechnung der Nährwert-Ähnlichkeit verwendet werden. Alle Daten und Diagramme können wie im Anhang beschrieben auf dem GitHub repository eingesehen werden. Die Abbildungen 11, 12, 13 und 14 zeigen auf der x-Achse jeweils die vier Benutzerprofile. Diese

unterscheiden sich durch die Anzahl an Rezepten. Auf der y-Achse findet man die Ergebniswerte der jeweiligen Methode für die verschiedenen Performanz-Kriterien.

Abbildung 11 Performanz-Kriterien Zufalls-Berechnung

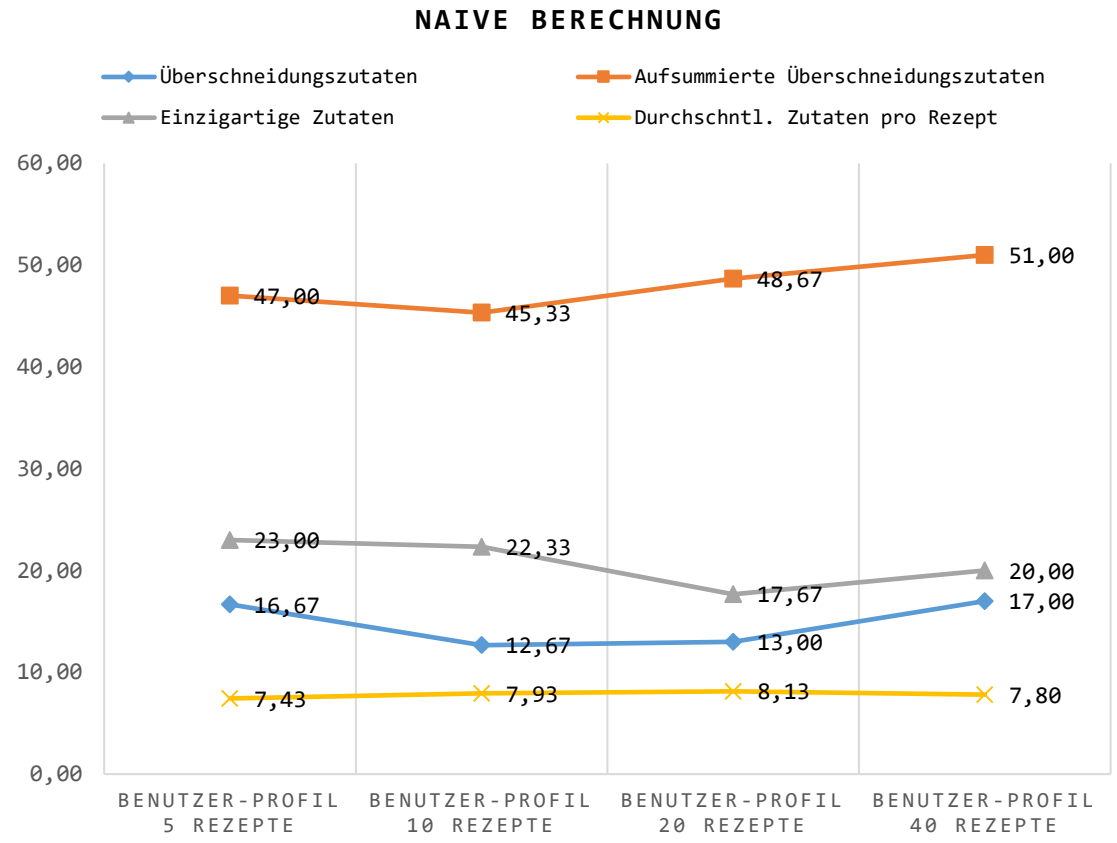


(Quelle: Eigene Darstellung und Berechnung)

Zu Beginn werden zunächst die Ergebnisse der Zufalls-Berechnung besprochen. Diese soll als Referenz dienen, um die drei Methoden besser einordnen zu können. In Abbildung 11 sind die Ergebniswerte die sich aus der Zufalls-Berechnung ergeben zu sehen. Was auffällt ist, dass beinahe lineare Anwachsen der Performanz-Kriterien *Überschneidungszutaten* und *aufsummierten Überschneidungszutaten*. Je mehr Rezepte im Benutzer-Profil sind, desto höher sind diese Werte. Die Wahrscheinlichkeit, dass Zutaten des Benutzer-Profils auch in der Top 10 Rezept-Liste sind, steigt mit der Anzahl der Rezepte im Benutzer-Profil. Die *einzigartigen Zutaten* der zufällig generierten Top 10 Rezepte liegen in einem hohen Wertebereich bei ungefähr 50.

Betrachtet man im Unterschied dazu das Diagramm für die Naive Berechnung (Abbildung 12), zeigt sich ein anderes Bild.

Abbildung 12 Performanz-Kriterien Naive Methode



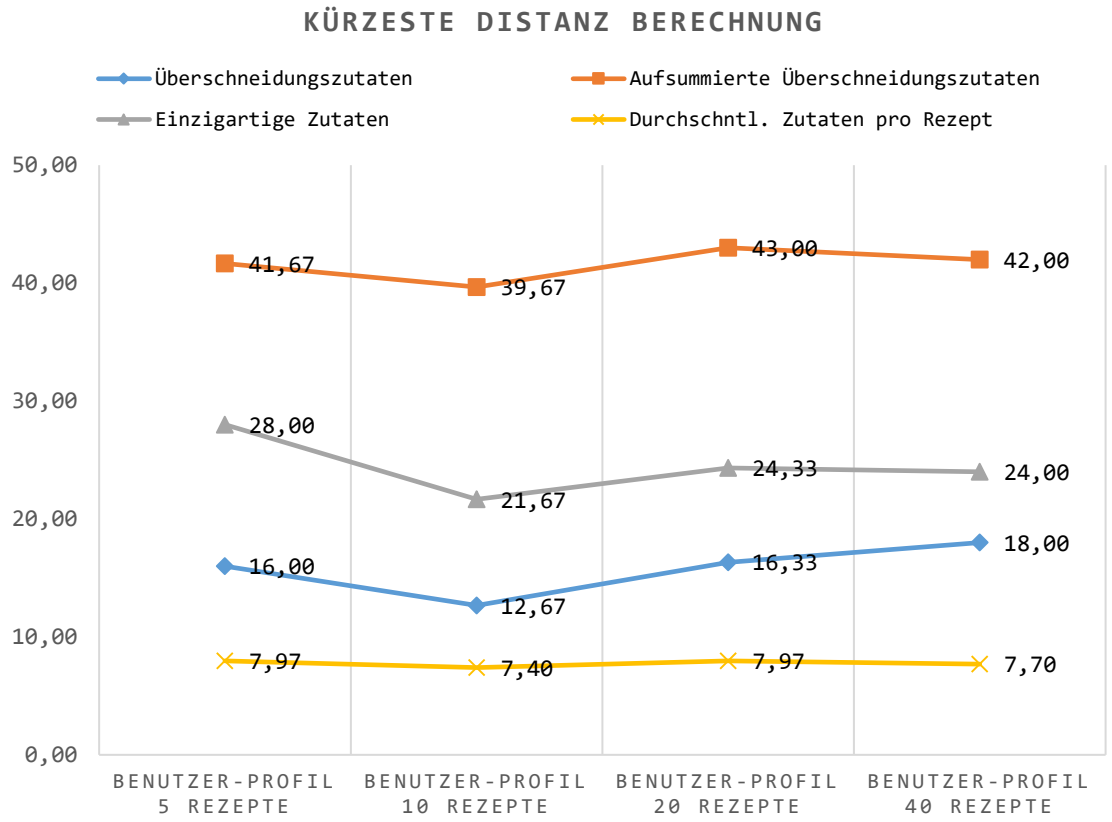
(Quelle: Eigene Darstellung und Berechnung)

Der Trend für die vier Performanz-Kriterien ist beinahe konstant. Die Anzahl der Rezepte im Benutzer-Profil hat fast keinen Einfluss auf die Werte. Schon bei fünf Rezepten im Benutzer-Profil sind die Werte auf einem hohen Niveau. Interessant ist die Entwicklung der *aufsummierten Überschneidungszutaten* und der *einzigartigen Zutaten*, im Vergleich zur Zufalls-Berechnung. Im Gegensatz zur Zufalls-Berechnung sind die Werte der *aufsummierten Überschneidungszutaten* beinahe konstant auf einem relativ hohen Niveau. Das bedeutet, dass die Naive Berechnung mehr Rezepte mit den gleichen *Überschneidungszutaten* empfiehlt. Auch sind die *einzigartigen Zutaten* im Gegensatz zur Zufalls-Berechnung auf einem viel niedrigeren Niveau. Das zeigt, dass eher Rezepte mit

gleichen Zutaten empfohlen werden. Die Empfehlungen der Naiven-methode sind so zusagen ‚spezifischer' was die Zutaten betrifft, wohingegen die Zufalls-Berechnung ‚breiter' streut, da ihre Empfehlungen zufällig sind.

Betrachtet man Abbildung 13, so stellt man fest, dass die Performanz-Kriterien der Kürzeste Distanz Berechnung sehr ähnliche Werte wie die der Naive Methode annehmen. Die Berechnungen der beiden Methoden zeigen keine signifikanten Unterschiede.

Abbildung 13 Performanz-Kriterien Kürzeste Distanz-Methode



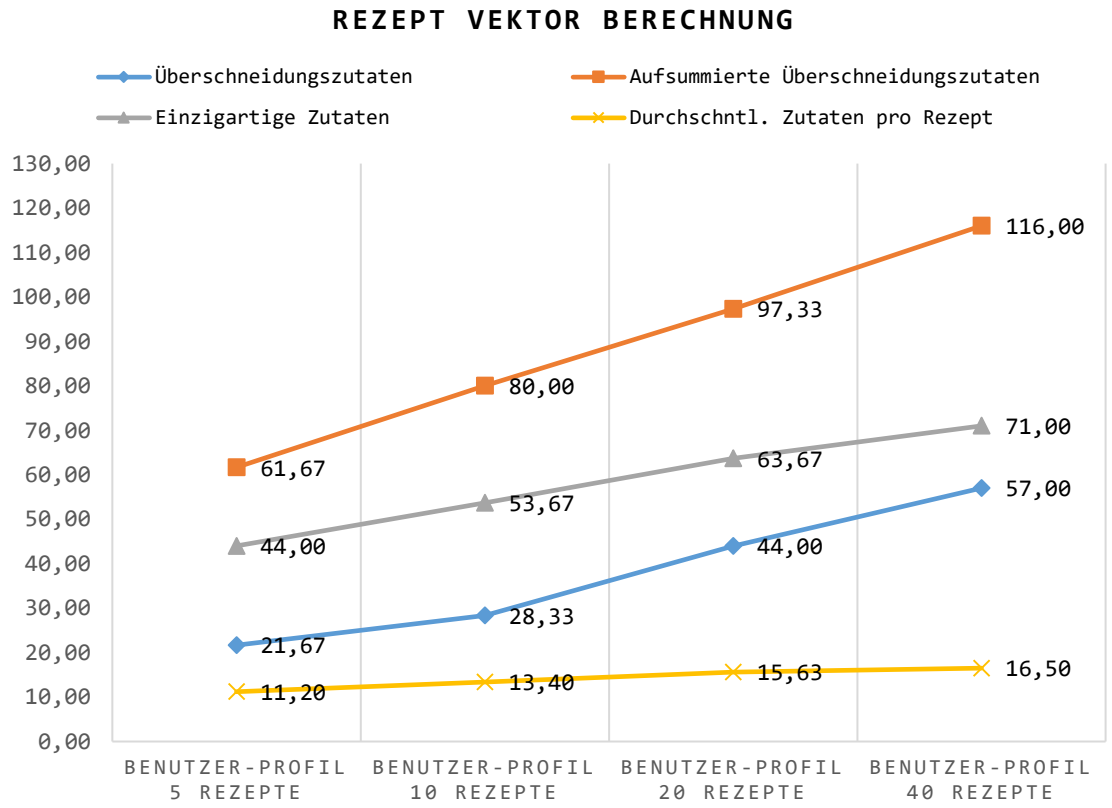
(Quelle: Eigene Darstellung und Berechnung)

Die Rezept Vektor-Methode ist die Methode, die sich von den anderen beiden Methoden am stärksten absetzt. Das liegt vor allem an ihrer Konzeption. Zwar verzeichnen die Performanz-Kriterien *Überschneidungszutaten*, *aufsummierte Überschneidungszutaten* und die *einzigartigen Zutaten* ein beinahe lineares Wachstum, abhängig von der Anzahl der Benutzer Zutaten, jedoch liegt das vor allem an der anwachsenden Anzahl an

Durchschnittszutaten von 11,2 Zutaten bei fünf Rezepten auf 16,50 Zutaten bei 40 Rezepten im Benutzer-Profil. Die *Durchschnittszutaten* der Methoden Naiv und Kürzeste Distanz liegen im Gegensatz dazu bei ungefähr acht Zutaten.

Die Rezept Vektor Methode ist so konzipiert, dass mit ansteigender Anzahl an Benutzer-Rezepten der Vektor immer mehr Zutaten in sich abbildet. Das heißt der Zutaten-Vektor ist ähnlicher zu den Rezepten, die auch viele Zutaten besitzen. Aus diesem Grund sind die Werte der Performanz-Kriterien – relativ zu den anderen Methoden – sehr hoch.

Abbildung 14 Performanz-Kriterien Rezept Vektor-Methode



(Quelle: Eigene Darstellung und Berechnung)

Wenn man jedoch das Verhältnis von *Überschneidungszutaten* zu *Durchschnittszutaten* für die drei hier besprochenen Methoden betrachtet, zeigt sich ein anderes Bild. Dieser Quotient liegt für 40 Benutzer-Rezepte bei der Vektor-Methode bei 3,45. Bei der Kürzeste Distanz Methode bei 2,34 und bei der Naive Methode bei 2,18. Das bedeutet, dass die

Rezept Vektor Methode trotz der konzeptionell hohen Anzahl an Zutaten in den vorgeschlagenen Rezepten, dennoch überdurchschnittlich viele *Überschneidungszutaten* in seinen Empfehlungen besitzt.

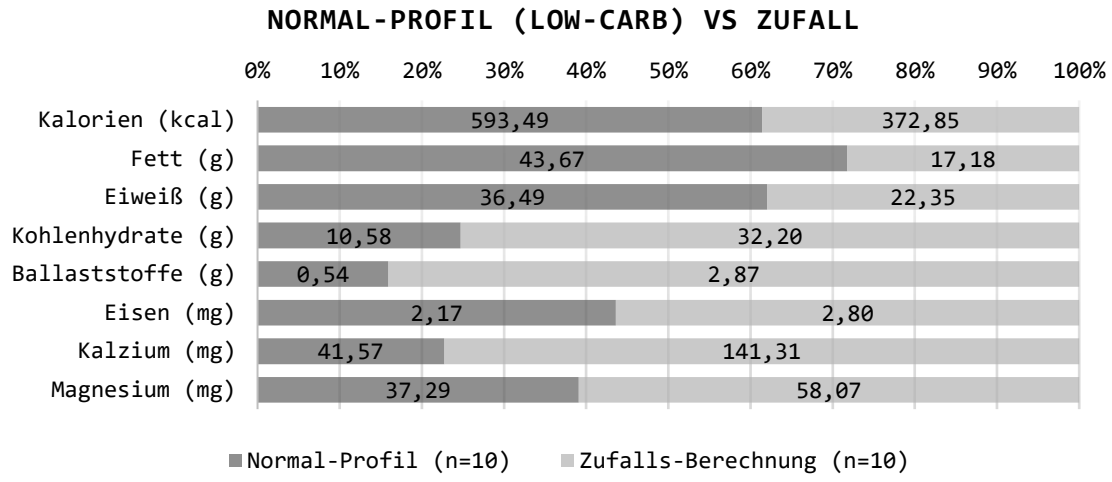
6.4 Ergebnisse Nährwert-Ähnlichkeit

Für die Berechnung der Nährwert-Ähnlichkeit für das Fett- und das Normal-Profil wurde die Naive Methode mit zehn Benutzer-Rezepten gewählt. Ihre Performanz-Kriterien unterscheiden sich nicht signifikant von der Kürzesten Distanz Methode. Die Rezept Vektor Methode wurde aufgrund ihrer Eigenschaft Rezepte mit überdurchschnittlich vielen Zutaten zu empfehlen, als nicht geeignet eingestuft.

Mit der Naive-Methode wurde sowohl für das Normal- als auch für das Fett-Profil eine Liste mit den Top 500 ähnlichsten (Jaccard Distanz) Rezepten berechnet. Zu dieser Liste wurde jeweils mit dem low-fat und dem low-carb Profilen die Euklidische Distanz berechnet.

Die Abbildungen 15 und 16 zeigen den Durchschnitt der Nährwerte aus den Top 10 ähnlichsten (Euklidische Distanz) Rezepten. Die Durchschnittswerte auf der rechten Seite der Diagramme zeigen die Nährwerte von zehn zufällig ausgesuchten Rezepten.

Abbildung 15 Nährwert-Ähnlichkeit Normal-Profil Durchschnitt



(Quelle: Eigene Darstellung und Berechnung)

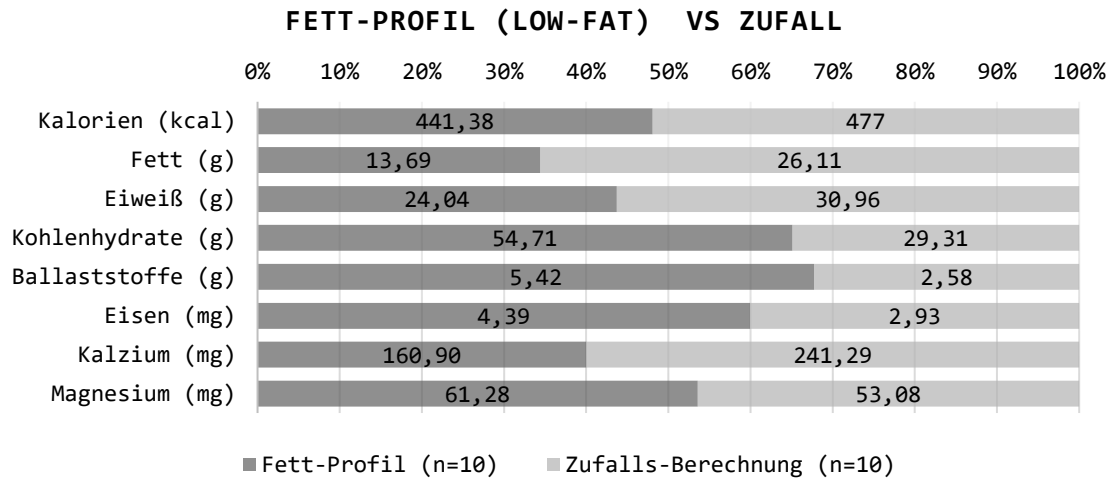
Die Nährwerte der Rezept-Empfehlungen des low-carb Profils für das Normal-Profil sind in Abbildung 15 dargestellt. Der wichtigste Wert ist der Wert für Kohlenhydrate. Dieser soll sehr niedrig sein und laut dem low-carb Profil den Wert 10g haben. Dieser Wert wird auch beinahe perfekt getroffen. Auch der Wert für den Fettgehalt der Rezepte liegt mit 43,67g liegt sehr nah an dem gewünschten Wert im low-carb Profil. Der Wert für Eiweiß liegt mit 36,49g niedriger als der angestrebte Wert von 53,3g. Die Nährwerte der zehn zufälligen Rezepte sind erwartungsgemäß sehr weite von den Ziel-Werten des low-carb Profils entfernt.

Tabelle 9 Auszug aus den Nährwerten des low-fat und low-carb Profils

	Kalorien	Fett	Eiweiß	Kohlenhydrate	Ballaststoffe	Eisen	Kalzium
low-fat	716 kcal	10 g	51,75 g	103,7 g	12,67 g	2 mg	266.67 mg
low-carb	716 kcal	41,3 g	53,3 g	10 g	12,67 g	2 mg	266.67 mg

(Quelle: vgl. NCBI, 2020 & vgl. Schnur, 2013, S.39)

Abbildung 16 Nährwert-Ähnlichkeit Fett-Profil Durchschnitt



(Quelle: Eigene Darstellung und Berechnung)

Die Mineralstoffe Kalzium und Magnesium weichen stark von den Zielwerten ab, da sie nicht gewichtet sind. Das bedeutet ihrer Ähnlichkeit zum low-carb Profil wird wenig Priorität zugeschrieben. Nur die Makronährstoffe sind gewichtet, mit dem größten Gewicht auf Fett (low-fat) und Kohlenhydrate (low-carb).

Die Nährwert-Ähnlichkeit für das Fett-Profil in Abbildung 16 zeigt ein ähnliches Bild wie das des Normal-Profils. Der entscheidende Makronährwert ‚Fett' ist mit 13.69g recht nah an den 10g des low-fat Profils. Jedoch werden die Zielwerte für die Nährstoffe Kohlenhydrate und Eiweiß nicht erreicht. Sie sind im Vergleich zu den Werten des low-fat Profils beinahe doppelt so niedrig. Insgesamt ist auch die Kalorienanzahl zu niedrig für ein Hauptgericht.

Tabelle 10 zeigt die Performanz-Kriterien der zehn Rezepte, die mit Hilfe der Nährwert-Ähnlichkeit empfohlen wurden. Vergleicht man die *Überschneidungszutaten* des Normal- und Fett-Profils mit den Berechnungen der Naive Methode für zehn Benutzer-Rezepte (Abbildung 12), so stellt man fest, dass sie mit 17 Zutaten recht hoch sind. Der hohe Wert lässt sich über die überdurchschnittlich hohe Anzahl an *einzigartigen Zutaten* (47 und 37) erklären. Denn je mehr einzigartige Zutaten in der Top 10 Empfehlung vorkommen, desto wahrscheinlicher ist es, dass dieselben Zutaten sich auch im Benutzer-Profil befinden. Die Werte für die *einzigartigen Zutaten* sind mit den Werten aus der Zufalls-Berechnung (Abbildung 11) vergleichbar. Das lässt sich dadurch erklären, dass eine gute Nährwert-Ähnlichkeit nur auf Kosten einer schlechteren Zutaten-Ähnlichkeit zustande kommt. Jedoch schneiden die beiden Profile bei den *aufsummierten Überschneidungszutaten* (34 und 30) deutlich besser als die Zufalls-Berechnung (16) ab.

Tabelle 10 Performanz-Kriterien Nährwert-Ähnlichkeit Normal- und Fett-Profil

Performanz-Kriterien	Normal-Profil	Fett-Profil
Überschneidungszutaten	17,00	17,00
Aufsummierte Überschneidungszutaten	34,00	30,00
Einzigartige Zutaten	47,00	37,00
Durchschntl. Zutaten pro Rezept	9,20	8,00

(Quelle: Eigene Darstellung und Berechnung)

7 Bewertung der Ergebnisse

Hypothese 1: *„Wenn ein Rezeptvorschlag mit der Jaccard Distanz generiert wird, dann hat er eine höhere Überschneidung an Zutaten zum Benutzter-Profil als eine zufällig generierter Rezeptvorschlag.“*

Betrachtet man nur die einzigartigen Überschneidungszutaten, dann schneiden die Berechnung des Zufalls mit 25,67 und des Rezept Vektors und 57 Zutaten bei 40 Benutzer-Profil Rezepten am besten ab. Jedoch relativiert sich dieses Ergebnis bei der Zufalls-Berechnung, wenn man die *aufsummierten Überschneidungszutaten* betrachtet. Diese liegen mit 32,67 bei 40 Benutzer-Profil Rezepten unter den Werten aller drei Zutaten-Ähnlichkeit Methoden bei nur 5 Benutzer-Profil Rezepten. Das bedeutet, dass die durch den Zufall gefunden Zutaten, im Gegensatz zu den drei vorgestellten Methoden, sich selten wiederholen.

Allgemein lässt sich sagen, dass Hypothese 1 bestätigt worden ist, denn die aufsummierten Überschneidungszutaten der Methoden Naiv, Kürzeste Distanz und Rezept Vektor sind immer höher als die des Zufalls.

Hypothese 2: „*Wenn eine Rezeptvorschlag mit der Euklidischen Distanz berechnet wird, dann sind die Nährwerte der Rezepte näher an den Nährwerten des low-fat bzw. low-carb Profils als ein zufällig generierter Rezeptvorschlag.“*

Man kann diese Hypothese als vollständig bestätigt sehen. Nährwerte der Top 10 empfohlenen Rezepte für das Normal- und das Fett-Profil sind sehr nah an den Werten der low-fat und low-carb Profile. Auch die Werte der Performanz-Kriterien sind besser als bei der Zufalls-Berechnung. Die Ergebnisse könnten wahrscheinlich noch besser ausfallen, wenn das in Kapitel 4.4 beschriebene Problem der Nährwert Normalisierung gelöst werden würde.

Hypothese 3: *„Je mehr Rezepte ein Benutzer in seinem Profil besitzt, desto höher ist die Überschneidung zwischen den Benutzter-Profil-Zutaten und den Rezeptvorschlags-Zutaten.“*

Die dritte Hypothese kann mit einem ‚ja‘ und einem ‚nein‘ beantwortet werden. Denn die Anzahl an Benutzer-Rezepten hat im Falle der Zufalls-Berechnung tatsächlich dazu geführt, dass die Anzahl der *Überschneidungszutaten* zugenommen hat. Das hat natürlich damit zu tun, dass durch das Erhöhen der Benutzer-Rezepte die Wahrscheinlichkeit steigt, dass Zutaten des Benutzer-Profils sich mit den Zutaten der zufällig erzeugten Top 10 überschneiden.

Für die drei Methoden der Zutaten-Ähnlichkeit trifft Hypothese 3 teilweise ein. Für die Methoden Naive- und Kürzeste Distanz-Berechnung trifft sie nicht ein. Denn die *Überschneidungszutaten* der beiden Methoden bleiben auch bei einer Erhöhung der Benutzer-Rezepte relativ konstant. Anders sieht es bei der Rezept Vektor Methode aus. Die *Überschneidungszutaten* wachsen hier – abhängig von der Anzahl der Benutzer-Rezepte – beinahe linear. Das liegt jedoch vor allem an der Anzahl der *durchschnittlichen Zutaten* der Rezept-Empfehlungen. Je mehr Zutaten in einem Rezept sind, desto höher ist auch die Wahrscheinlichkeit *Überschneidungszutaten* zum Benutzer-Profil zu erhalten.

8 Fazit und Ausblick

Recommender System zu erstellen, die den ‚echten' Geschmack eines Nutzers abbilden ist beinahe ein philosophisches Problem, denn man ist mit der Frage konfrontiert: „Was macht Rezepte und die damit verbundene Geschmackserlebnisse aus?" Wie misst man Geschmack und die Präferenzen von Menschen? Eins ist sicher, ein zu einfaches Modell, das nur über Zutaten funktioniert kommt der Lösung zu diesem Problem nicht richtig nahe. Vielversprechender sind Wissenschaftler wie Ahn, Ahnert, Bagrow und Barabási, die sich dem Problem über das Messen von ‚echten' Geschmacksrichtungen und ihren Kombinationen näheren.

Ziel dieser Arbeit war es jedoch nicht, wissenschaftliche Sensation zu schaffen. Vielmehr sollte ein gesamter data-science Prozess von Anfang bis zum Schluss abgebildet und erlebt werden. Vom ersten Schritt der Daten-Aquise mit Hilfe eines Bot-Programms, über die Datenmanipulation mit unterschiedlichen Programmbibliotheken bishin zur Auswertung und Interpretation der Ergebnisse.

Zusammenfassend kann man sagen, dass die Bestimmung von Zutaten-Ähnlichkeiten mithilfe von Abstandsmaßen funktionieren kann. Die drei hier vorgestellten Mehtoden haben ähnliche Rezepte gefunden. Vor allem die Rezept-Vektor Methode hat im Verhältnis zwischen *Überschneidungszutaten* und *durchschnitlliche Zutaten pro Rezept Top 10* am besten abgeschnitten. Man müsste überprüfen, welche Ergebnisse diese Methode liefert, wenn man die Anzahl der Zutaten anpasst.

Zudem muss hinterfragt werden, ob es nicht besser Performanz-Kriterien gibt, als die die in dieser Arbeit entwickelt wurden.

Die in dieser Arbeit vorgestellten Methoden weisen verlässliche Ergebnisse bezüglich der Nährwert-Ähnlichkeit auf. Deshalb könnten sie in Zukunft vielleicht in der Diätplanung Verwendung finden.

9 Quellenverzeichnis

Ahn, Y.Y.; Ahnert, S.E.; Bagrow, J.P. & Barabási, A.L. (2011): Flavor network and the principles of food pairing, in: Scientific Reports, S. 1-7

Allrecipes.com (2015): Facts & Stats, Internet: https://web.archive.org/web/20150321094241/http:/press.allrecipes.com/fact-stats/, Zugriff: 16.05.2020

Blitzstein J.; Pfister H. (2013): CS 109/Stat 121/AC 209/E-109 at Harvard, Internet: https://www.quora.com/What-is-it-like-to-design-a-data-science-class-In-particular-what-was-it-like-to-design-Harvards-new-data-science-class-taught-by-professors-Joe-Blitzstein-and-Hanspeter-Pfister, Zugriff: 15.05.2020

Cromwell, E.; Galeota-Sprung. J.; Ramanujan, R. (2015): Computational Creativity in the Culinary Arts, in; Proceedings of the 46th ACM Technical Symposium on Computer Science Education - SIGCSE ’15, S. 38-42

Devaux, M.; Vuik, S. (2019): The Heavy Burden of Obesity, in: OECD Health Policy Studies, vol. 19, Paris: OECD Publishing

FAO (2006): Livestock’s long shadow - environmental issues and options

Freyne, J.; Berkovsky, S. (2010): Recommending food: Reasoning on recipes and ingredients, in: Lecture Notes in Computer Science, 6075 LNCS, Berlin, Heidelberg: Springer, S. 381-386

Huang, A. (2008): Similarity measures for text document clustering, in: New Zealand Computer Science Research Student Conference, NZCSRSC 2008 - Proceedings, S. 49-56

NCBI (2020): Recommended Dietary Allowances and Adequate Intakes, Total Water and Macronutrients, Dietary Reference Intakes, Internet: https://www.ncbi.nlm.nih.gov/books/NBK56068/table/summarytables.t4/?report=objectonly, Zugriff: 12.05.2020

NCBI (2020): Recommended Dietary Allowances and Adequate Intakes, Total Water and Macronutrients, Dietary Reference Intakes, Internet: https://www.ncbi.nlm.nih.gov/books/NBK56068/table/summarytables.t5/?report=objectonly, Zugriff: 12.05.2020

Niwattanakul, S.; Singthongchai, J.; Naenudorn, E.; Wanapu, S. (2013): Using of jaccard coefficient for keywords similarity, in: Lecture Notes in Engineering and Computer Science, S. 380-384

Pazzani, M.J. & Billsus, D., (2007): Content-based recommendation systems, Lecture Notes in Computer Science, 4321 LNCS, S. 325-341

Portugal, I.; Alencar, P.; Cowan, D. (2015): The Use Of Machine Learning Algorithms In Recommender Systems: A Systematic Review, S. 1-16

Schnur, E. (2013): Umsetzung der D-A-CH-Referenzwerte in die Gemeinschaftsverpflegung, Bonn: Deutsche Gesellschaft für Ernährung e. V.

Teng, C.-Y.; Lin, Y.-R.; Adamic, L.A. (2011): Recipe recommendation using ingredient networks, in: Proceedings of the 4th Annual ACM Web Science Conference, WebSci'12, S. 298-307

10 Anhang

Repository aller Daten und Berechnung, die in dieser Arbeit verwendet wurden:
https://github.com/DanielVolz/ba_recommender_system_2020

Repository des Bot-Programms, zum Herunterladen der Rezept-Daten:
https://github.com/DanielVolz/allrecipes_crawler

Hinweis: Zum Ausführen der Ergebnis Skripte im *.js Format ist neben der Datenbankverbindung zu den hochgeladenen Rezept-Daten das Programm Robo 3T (https://robomongo.org/) notwendig/empfohlen.